```
RESULT  9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;
```

```
  Query Match             7.3%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 3.7e-14;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy    313 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 371
          ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db    599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy    372 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 418
           :  |  |   |  ::  | |:|| :  | :||  |        ||:|| : :    |:||
Db    652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy    419 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 477
           : :   ||  || ::    |:|| : :       | :||  || ::      |     :
Db    712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy    478 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 533
            ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db    765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy    534 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 588
            ||    | :||  |  ::       |      | : |  || ::   | |:||       ||
Db    818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy    589 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 638
          ·||:  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db    877 ----PSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

                                  APTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 698
                                  | ::   | |:||  ::       | :||  |       |
                                  ADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

                                  GTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 757
                                    |:|| :  | :||  || ::   |  | :   | :||
                                  ---PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

                                 FSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 813
                                  :|  | :||       :| :||  || ::    |  || ||
                                 PADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy                               THKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 872
                                  | :|  ||| : ||  ||                 ||| :
Db   1085 DSSAH-STPSTPAD--AHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy    873 MT 874
          |
Db   1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
```

```
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;
```

```
  Query Match             7.3%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 3.1e-14;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy    284 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 339
          :| |  :| | ||    |      ||   |.     |  | : ||  | ||     | |
Db      8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy    340 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 394
            |||    :|  ||    ||     | |     || :    |           |||  |
Db     62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy    395 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 440
            |  ||    |       | |   :  ||   ||       : |  || ||   |
Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    441 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 492
            ||  |      ||      ||   : |||    | |  | |:  ||     :   :
Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    493 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 548
          |    ||   ||  |::    |   ||  :  ||  |     |   : : |         ||
Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    549 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 589
           ||         : |  |      ||   ||   ||    ||         | |||  : |
Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    590 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 645
            | : ||     | | |   || | |:      || |   ||:   :|   || : ||
```

```
Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    646 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 694
                |    || |     |  | ||   :|| |:             ||:     ||
Db    414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    695 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 746
            ||  :  |  ||:  |  | |     || |        | |  ||  |    ||    |
Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    747 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 792
            | |        |  |  | |  |  ||:    :   ||:|  | | |    | |   :
Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    793 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 842
              || |  :| |     | |: :  : |   : |. | :: |           | ||
Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    843 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 900
          : |  |:|:     :| :  |: | |:        |     :: |||  |  ||| :::
Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    901 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 953
                :     :|       | |          | | | :    : :  | | ::
Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    954 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 1013
                : |: |:        : |  |  : ||   : |  :  |     |   | ||
Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy   1014 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 1056
            | : :|   |   :   :|:   |:      ||    :| |:      |||
Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864
```

```
RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
```

```
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN       21     41       NONREPETIVE LINKER.
FT   DOMAIN      124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                                P-[PST]-[ST].
FT   DOMAIN      184    192       NONAPEPTIDE 1.
FT   DOMAIN      213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;


  Query Match             7.1%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity   29.2%;  Pred. No. 8.5e-14;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy    180 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 230
          |   | ||   | :| | | :    ::|       ||: | | |  :|   | |  |
Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy    231 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 290
          :|    |   :|          | :|    : | | : ::    | |      |:
Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    291 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 349
          || |       | :| |  ||    |  | |  ||   | | |     ||   | | |
Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    350 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 404
          |: || :|  |   | :|  | |    ||  ::|: |     |  ||   | |  ||
Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    405 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 457
          |:  |  |        |:  | |  ||   | | :| |: |    ||  ||   | |
Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    458 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 512
          | | |: |    |  | |:   ||   ||  |  |:       :|   ||: ||  | |
Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    513 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 562
            |  |  |     |  |    |||  :||  ||        :| |: |    | ||  |
Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    563 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 617
            ||   | | |  |         : |   ||   |::     || |: |    |
Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    618 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 668
          ||    |  |    ||   | | | :|:  || :  |: |      |   ||:   |:| |
Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    669 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 725
            |  || | | |  |      ||   |    | ||  ||      :  ||  || : | | |
Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    726 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 781
                      || | | | | |: |    :||:   :  : :  ||:  | |
Db    584 -------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    782 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 830
          | |  || | | | | | |: |   :|  | | |       |    |: | | : :   |
Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    831 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 886
            |    |: :  || |     ||    || |    |||    : | ||    | :
Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    887 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 946
            : | |  | |     :   | |   |  ::   ::    |    |   :: |     ||
Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    947 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 1001
          :                :   | |||     | |:: ||:    ||  ||    |: ||| |
Db    786 I--------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy   1002 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 1043
            |   ||| :  || :   :  |   |       |     : |:|
Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871


RESULT  12
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
```

```
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;


  Query Match             7.1%;  Score 517.5;  DB 1;  Length 831;
  Best Local Similarity   26.6%;  Pred. No. 8.9e-14;
  Matches  217;  Conservative   76;  Mismatches  345;  Indels  179;  Gaps   41;

Qy    105 KKTKKVIESE--EITEEHSVS-ENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLK 161
          | ||: :| :  |: : | |   : : :     :   :| |::     :|  || |  |
Db    103 KSTKESLERQRSELEDRHQVDMASYQDAIQQLDNELRNTKWEM-----AAQLREYQDLLN 157

Qy    162 VK--------------DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 207
          ||              : :: |    | |           |  |  |:  |  ||  | |
Db    158 VKMALDLEIAAYRKLLEGEECRIGFGPIP---------FSLTEGLPKI--PSMST--HIK 204

Qy    208 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTT 267
          | :  ||   :             : ||| :   :   ::  |  |  :     |:|:
Db    205 VKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEEDKEAQGEEEEE 252

Qy    268 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----TPTTPKEPA-S 322
          : :  :    ||     | ::  | |    |  || |     | |    :|   | || :
Db    253 AEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKSPAEAKSPAEA 303

Qy    323 TTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTKEP----APTT 370
           :| |     :  :|   | ||    |   || |   | ||    | | | |    :|
Db    304 KSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVKSPGEAKSPAE 362

Qy    371 PKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA----PTTPKEPA 414
          | ||    |   | ||    | : || |   | ||    |   | ||    |   | ||
Db    363 AKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATVKSPVEAKSPA 421

Qy    415 ----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTAPKKPAPTTPK 458
              | | | |    :|   | || : | |    :|   | ||    |   | ||    |
Db    422 EVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVVAKSPAEA--K 478

Qy    459 EPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEPAPTTTKS--- 506
          ||   |  | : :  ||   | ||    |  ||    :|   | |     ||
Db    479 SPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSPVKEGAKSLAE 538

Qy    507 --APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTK 560
            :|   | |     | ||    |   | |     | |     :      | ||    |
Db    539 AKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPEAKPPAKEEAK 598

Qy    561 KPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTP 620
          :||     |   :  || | :  |: | |  ||:||   |  :|   ||:     :||
Db    599 RPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV---KAKEPPK 651

Qy    621 TTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLK 680
            ||  | |||     : |: ||   ::|     |||    ||::     |  |    |
Db    652 KVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEAKKEEA----K 706

Qy    681 EPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTP 740
          |     |::  | :|      :|     |  |             |||:  : ||     |
Db    707 EKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKPSEKE----KP 750

Qy    741 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTT 800
          |       ||  |  |:|   ||          ||  |     ||  |    ::
Db    751 K-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAKAKEED----- 789

Qy    801 PKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 837
            |  |  |  |    :| |: |  |:     |:|::
Db    790 --KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824


RESULT  13
MUC1_XENLA
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
```

```
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                 ALA/THR-RICH.
FT   REPEAT        81     88       1-1.
FT   REPEAT        89     96       1-2.
FT   REPEAT        97    104       1-3.
FT   REPEAT       105    112       1-4.
FT   REPEAT       113    120       1-5.
FT   REPEAT       121    128       1-6.
FT   REPEAT       129    136       1-7.
FT   REPEAT       137    144       1-8.
FT   DOMAIN       161    202       P-TYPE 1.
FT   DOMAIN       218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT       218    224       2-1.
FT   REPEAT       225    239       2-2.
FT   REPEAT       240    249       2-3.
FT   REPEAT       250    259       2-4.
FT   REPEAT       260    275       2-5.
FT   REPEAT       276    287       2-6.
FT   REPEAT       288    294       2-7.
FT   REPEAT       295    301       2-8.
FT   DOMAIN       306    347       P-TYPE 2.
FT   DOMAIN       353    394       P-TYPE 3.
FT   DOMAIN       402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                 THR-RICH.
FT   REPEAT       402    411       3-1.
FT   REPEAT       412    419       3-2.
FT   REPEAT       420    431       3-3.
FT   REPEAT       432    443       3-4.
FT   REPEAT       444    453       3-5.
FT   REPEAT       454    460       3-6.
FT   REPEAT       461    472       3-7.
FT   REPEAT       473    479       3-8.
FT   REPEAT       480    491       3-9.
FT   REPEAT       492    498       3-10.
FT   REPEAT       499    515       3-11.
FT   REPEAT       516    522       3-12.
FT   DOMAIN       525    566       P-TYPE 4.
FT   DOMAIN       572    613       P-TYPE 5.
FT   DOMAIN       620    661       P-TYPE 6.
FT   DISULFID     162    188       BY SIMILARITY.
FT   DISULFID     172    187       BY SIMILARITY.
FT   DISULFID     182    199       BY SIMILARITY.
FT   DISULFID     307    333       BY SIMILARITY.
FT   DISULFID     317    332       BY SIMILARITY.
FT   DISULFID     327    344       BY SIMILARITY.
FT   DISULFID     354    380       BY SIMILARITY.
FT   DISULFID     364    379       BY SIMILARITY.
FT   DISULFID     374    391       BY SIMILARITY.
FT   DISULFID     526    552       BY SIMILARITY.
FT   DISULFID     536    551       BY SIMILARITY.
FT   DISULFID     546    563       BY SIMILARITY.
FT   DISULFID     573    599       BY SIMILARITY.
FT   DISULFID     583    598       BY SIMILARITY.
FT   DISULFID     593    610       BY SIMILARITY.
FT   DISULFID     621    647       BY SIMILARITY.
FT   DISULFID     631    646       BY SIMILARITY.
FT   DISULFID     641    658       BY SIMILARITY.
FT   VARIANT      276    276       K -> E.
FT   VARIANT      354    354       C -> R.
FT   VARIANT      415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;

  Query Match             7.0%;  Score 509.5;  DB 1;  Length 662;
  Best Local Similarity   28.1%;  Pred. No. 1.5e-13;
  Matches  223;  Conservative   48;  Mismatches  276;  Indels  247;  Gaps   30;
```

```
Qy    253 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 312
          ||    :  ||: | :|: : : |||:|       :| | ||  |     |   ||
Db      3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy    313 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 372
                |: |    |      ||||   || |    ||||    ||||    ||||
Db     52 --------AAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy    373 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 430
          || |   | |||  | ||||      :|    ||||    | :  :  : :: :|
Db    104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy    431 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 475
            |: :|   :   | |:      ||     ||          :|   : :|   ||||
Db    164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy    476 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 535
                |||| : | |||    ||||           |||  :  |||
Db    219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy    536 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 595
                    ||  :  ||||             || | |  ||    ||||
Db    246 ------------TTTTKATPTTT------------TTTKATTTTT----TPTT----- 270

Qy    596 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 655
                ||    | |||   | |||
Db    271 ---------TTTTTKATTTPTTTTTTTP------------------------------- 289

Qy    656 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 715
                      |  ||   |  ||       | |              :|   :    :
Db    290 ----------TTTTTKATTTTTTTSGECKMEPSK-----------REDCGYSGITESQCR 328

Qy    716 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 765
          ||     : | |              ||:     |    |    |    :|    : :
Db    329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy    766 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 825
          : ||    :||   |  |   | |||   |  ||  |   |||       |:  :|||||
Db    385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTPTTT 436

Qy    826 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 885
          ||                    | ||              || ||  | |  |||   | |
Db    437 TTTTT----------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy    886 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 945
          |||  |||  |     | |||: |        |||  |:|||  ||  | ||   ||
Db    464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy    946 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 1001
          ||||  | |||    | |    |  ||        : |  :|     :|  : |    :
Db    511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy   1002 MPRVRKPKTTPTPR 1015
          :|:|   |  |: |
Db    567 LPQVADCKVAPSSR 580
```

```
RESULT  14
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
```

```
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;


  Query Match             6.9%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 6.4e-13;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     312 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 365
           |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     366 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 423
           | | :|  | |::   |:|    |  :|  ||   |: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     424 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 483
           |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     484 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 543
           :||:   :||:   :||:   :||:|   |||:   :||:|   :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     544 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 603
              :||:|   :||:    :||:|   :||:|   | ||:|   :||:|   :||:|
Db    1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     604 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 661
           ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     662 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 720
           :|   :||:|| |          ||:|   |:  |  ||: |   | || | :||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy     721 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 779
           ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     780 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 837
           :||:|   :||:||    :||:|   :||:|      |  :::|          |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     838 STPE 841
           |  |
Db    1966 SDEE 1969


RESULT  15
NFH_HUMAN
ID   NFH_HUMAN      STANDARD;      PRT;  1020 AA.
AC   P12036;
DT   01-OCT-1989 (Rel. 12, Created)
DT   01-OCT-1989 (Rel. 12, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=88328981; PubMed=3138108;
RA   Lees J.F., Shneidman P.S., Skuntz S.F., Carden M.J., Lazzarini R.A.;
RT   "The structure and organization of the human heavy neurofilament
RT   subunit (NF-H) and the gene encoding it.";
RL   EMBO J. 7:1947-1955(1988).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
```

```
DR   EMBL; X15306; CAA33366.1; -.
DR   EMBL; X15307; CAA33366.1; JOINED.
DR   EMBL; X15308; CAA33366.1; JOINED.
DR   EMBL; X15309; CAA33366.1; JOINED.
DR   PIR; S00979; QFHUH.
DR   MIM; 162230; -.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation.
FT   DOMAIN        1    100       HEAD.
FT   DOMAIN      101    413       ROD.
FT   DOMAIN      414   1020       TAIL.
FT   DOMAIN      101    132       COIL 1A.
FT   DOMAIN      133    145       LINKER 1.
FT   DOMAIN      146    244       COIL 1B.
FT   DOMAIN      245    266       LINKER 12.
FT   DOMAIN      267    288       COIL 2A.
FT   DOMAIN      289    292       LINKER 2.
FT   DOMAIN      293    413       COIL 2B.
SQ   SEQUENCE   1020 AA;  111780 MW;  1177C9DCB3DCF1D4 CRC64;


  Query Match             6.9%;  Score 502;  DB 1;  Length 1020;
  Best Local Similarity   26.7%;  Pred. No. 4.2e-13;
  Matches  209;  Conservative   81;  Mismatches  356;  Indels  136;  Gaps   35;

Qy    105 KKTKKVIESE--EITEEHSVS-ENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLK 161
          | ||  :| :  |: : |     : : :     :   :| |::     :|  || |  |
Db    336 KSTKDSLERQRSELEDRHQADIASYQEAIQQLDAELRNTKWEM-----AAQLREYQDLLN 390

Qy    162 VK--------------DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 207
          ||              : :: |    | |           |  |  |:  |  ||  | |
Db    391 VKMALDIEIAAYRKLLEGEECRIGFGPIP---------FSLPEGLPKI--PSVST--HIK 437

Qy    208 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKE-KT 266
          | :  ||   :             : ||| :        :|:||  | : |  : || |
Db    438 VKSEEKIKVVE-----------KSEKETVIVEE-----QTEETQVTEEVTEEEEKEAKE 480

Qy    267 TSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPK 326
             || :  |:  |:              |  || |    | |   :  ||  |   :
Db    481 EEGKEEEGGEEEEAEG-------------GEEETKSP----PAEEAASPEKEAKSPVKE 522

Qy    327 EPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTT 386
          |        :|   :  :|   || |   | ||    || | : |:  :|  :  :|
Db    523 EAKSPAEAKSPEKEEAKSPAEVKS-PEKAKSPA----KEEAKSPPEAKSPEKEEAKSPAE 577

Qy    387 TKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAP 446
           || |   | ||    | ||    :  :|   | |     | ||   | |     | ||
Db    578 VKS-PEKAKSPAKEEAKSPA----EAKSPEKAKSPVKEEAKSPA-EAKSPVKEEAKSPAE 631

Qy    447 TAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 506
          : | :  || | :  |  :|   :  ||   | |     || |   | |     ||
Db    632 VKSPEKAKSPTKEEAKSPEKAKSPEKEEAKSPEKAKSPVKAEAKS-PEKAKSPVKAEAKS 690

Qy    507 APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTA 566
           |   | |     || | :  |  :|   :  :|   | |     || | |  |  :|
Db    691 -PEKAKSP----VKEEAKSPEKAKSPVKEEAKSPEKAKSPV----KEEAKTPEKAKSPV- 740

Qy    567 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPE--KPAPTTPEEL-APTTPEEPTPTTP 623
           || | :  |  :|   | |   :||   |   |   ||   ||:  :|   |  :|
Db    741 -KEEAKSPEKAKSPEKAKTLDVKSPEAKTPAKEEARSPADKFPEKAKSPVKEEVKSPEKA 799

Qy    624 EEPAPTTPKAAAPNTPKEPAPTTP----KEPAPTTPKEPAPTTPKETAPTTPKGTAPTTL 679
          : |     ||     ||:    :|    ::|     |||     :| || |||
Db    800 KSPLKADAKAPEKEIPKKEEVKSPVKEEEKPQEVKVKEPPKKAEEEKAPATPK------- 852

Qy    680 KEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 739
           |    : |: |||: ||    :|      :||   || :     || |    | |   |
Db    853 TEEKKDSKKEEAPKKEAP----KPKVEEKKEPAVEKPKESKVEAKKEEAEDKKKVP---T 905

Qy    740 PKGTAP--TTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPA-PTTPKETAPTTPKEP 796
          |:  ||    :|| |       |||      | |    : :|: |   || ||   |:
```

```
Db    906 PEKEAPAKVEVKEDA-------KPKEKTEVAKKEPDDAKAKEPSKPAEKKEAAP-EKKDT 957

Qy    797 APTTPKKPAPTTPETPPPTTSEV--STPTTTKEPTTIHKSPDE---STPELSAEPTPKAL 851
             ||     ||  | | ::      | :|||:       |   || :  ::|  ||
Db    958 KEEKAKK-----PEEKPKTEAKAKEDDKTLSKEPSKPKAEKAEKSSSTDQKDSKPPEKAT 1012

Qy    852 EN 853
          |:
Db   1013 ED 1014


Search completed: April 26, 2002, 16:32:55
Job time: 605 sec
```

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:25:49 ; Search time 61.21 Seconds
(without alignments)
502.197 Million cell updates/sec

Title: AA6
Perfect score: 7294
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1366

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	7.2	805	4	US-09-103-429A-4	Sequence 4, Appli
2	513	7.0	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	6.9	744	6	5202236-25	Patent No. 5202236
4	499.5	6.8	1721	3	US-08-700-651-5	Sequence 5, Appli
5	498.5	6.8	786	4	US-09-103-429A-3	Sequence 3, Appli
6	498.5	6.8	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	6.7	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	6.7	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	6.5	652	6	5202236-13	Patent No. 5202236
10	452	6.2	960	4	US-09-219-849-5	Sequence 5, Appli
11	425.5	5.8	1185	4	US-09-041-886-23	Sequence 23, Appl
12	424.5	5.8	1867	2	US-08-479-537A-5	Sequence 5, Appli
13	424.5	5.8	1867	4	US-09-083-116-5	Sequence 5, Appli
14	424.5	5.8	2035	2	US-08-479-537A-2	Sequence 2, Appli
15	424.5	5.8	2035	4	US-09-083-116-2	Sequence 2, Appli
16	419.5	5.8	2476	2	US-08-276-967-2	Sequence 2, Appli
17	417	5.7	829	1	US-08-642-255-132	Sequence 132, App
18	417	5.7	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	5.7	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	5.7	837	1	US-08-477-509B-103	Sequence 103, App
21	417	5.7	837	1	US-08-642-255-101	Sequence 101, App
22	417	5.7	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	5.7	837	3	US-08-482-085B-103	Sequence 103, App
24	417	5.7	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	5.7	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	5.7	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	5.5	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	5.5	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	5.5	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	5.5	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	5.5	682	1	US-08-397-633A-36	Sequence 36, Appl
32	387	5.3	1537	1	US-08-325-267A-2	Sequence 2, Appli
33	383.5	5.3	1231	3	US-08-904-263A-4	Sequence 4, Appli
34	381	5.2	1848	4	US-08-296-791-6	Sequence 6, Appli
35	381	5.2	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
36	376	5.2	960	4	US-09-219-849-6	Sequence 6, Appli
37	373	5.1	1786	4	US-08-973-462-8	Sequence 8, Appli
38	371	5.1	761	2	US-08-707-237A-84	Sequence 84, Appl
39	371	5.1	762	1	US-08-642-255-114	Sequence 114, App
40	371	5.1	762	1	US-08-397-633A-26	Sequence 26, Appl
41	370.5	5.1	1064	1	US-08-642-255-62	Sequence 62, Appl
42	370.5	5.1	1187	1	US-08-320-559-28	Sequence 28, Appl
43	370.5	5.1	1187	3	US-08-545-860D-28	Sequence 28, Appl
44	370.5	5.1	1187	5	PCT-US94-04496-28	Sequence 28, Appl
45	370.5	5.1	1210	1	US-08-320-559-26	Sequence 26, Appl

ALIGNMENTS

```
RESULT   1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-4


  Query Match             7.2%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity   27.8%;  Pred. No. 5.8e-26;
```

```
  Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy    286 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 340
          |: |    |:     | |||  |    |  | |     | |   |           |
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    341 KEPAPTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 388
          :: || |           :  | |  || ||    || | :  |||||:  |||||
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    389 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 448
           ||||  :   ||           ||||  : || |     | ||: |  ||
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    449 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 508
              ||||  : ||||  : |||||:          |||||: |||||:  ||||   |
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    509 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 552
          |||   :|  |  || |||    || |||  ||||:    ||  |
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    553 ------------------------------------------------EPAPTTTK 560
                                                           |||  |:
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    561 K----------------PA---------------------------------------- 563
                           ||
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    564 --------------------------------PTAP--KEPAPT-----------TPKET 578
                                           || |   || |           | |:
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    579 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 604
                                  :| |    :| |          :   |     ||
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    605 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 664
          |     |||      |:|   || | |  |||  |    ||    ||||     ||||
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    665 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 721
            :|||   |  |||    ||||    |  |:  | |  ||:  :   |||     ||
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    722 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 780
          ||     |||      |:|   ||||   ||| |          | |  ||:  |:  |
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    781 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 840
          ||     |  | |   ||||   ||| |  | ||| :  : |     ||  |      |
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    841 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 872
            :| |             | ||  || | ||
Db    686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             7.0%;  Score 513;  DB 3;  Length 1837;
  Best Local Similarity   21.6%;  Pred. No. 5.9e-25;
  Matches  307;  Conservative  105;  Mismatches  539;  Indels  470;  Gaps   49;

Qy     70 HNPTSPPSSKKAPP----PSGASQTIKSTTKRSPKP--------------PNKKKTKKV 110
          |   |  ||  : |     ||    |:  |:|   |               ||   | :
Db     83 HADASTTSSAYSAPFELDVSGV--PIEPNTRRMVDPVSLMLFDNSTGVMYDPN---TNSI 137

Qy    111 IESE--EITEEHSVSENQESSSSSSSSSSSSTIWKIKSSKNSAANRELQKKLKVKDNKKN 168
          :|     |  |  :      :|::  :: :|  | :  |  |   ::  |:  :  ::
Db    138 LEGSIAGIRSESCIVSELNFTSTTGFTTDTSMNWPV--SITSGELKDPNKQATISGSRSC 195

Qy    169 RTKKKPTPKPPVVDEAGSGLDNG-DFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLP 227
            |:            |  :|:   |:| :                   |  | :| |: |
Db    196 GWKQ-----------GYSIDSSTGFRVDS----------------ITGLPTDPYPNCP 226

Qy    228 PNSDTSKETSLTVNKET------TVETKETTTTNKQTST-----DGK-----EKTTSAKE 271
            |  |    | :  |           : || ||    :|     | |         | ::
Db    227 FNPVTGNLVSRSTGKTIPNTYAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQ 286

Qy    272 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPT 331
          |  :  |   :  |  |  | |    |||      ||    | ||     :||    | |
Db    287 VQIFDMGSKVYIPYTKCVGVKHTTTTTTTT-----TTTTTTTTTTTTTTTTTTTTTTTTT 341

Qy    332 TIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP 391
          |  :  ||      ||| :  ||      |||      || |:|  |||     ||| :
Db    342 TTTTTTTT------TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTT 395

Qy    392 TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKK 451
          ||      ||      ||      || |  | ||      ||      ||      |  ||
Db    396 TTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTKK 455

Qy    452 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTP 511
          |  ||      ||      |||   : ||      ||| :| ||||:|  |||  :  ||
Db    456 PTTTT------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATTTTKKPTTTTTTTTTTT 509

Qy    512 KEPSPTTTKEPAPTTPKEPAPTTPKK---------------------------------- 537
          |:|: |||     || :  :   | :
Db    510 KKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFT 569
```

```
Qy    538 ------------------------------------------------------------ 537

Db    570 MIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVS 629

Qy    538 ------------------PAPTTPKE------------------PAPTTPKEPAPTTTKK 561
                             ||  | |                    |  :   ||| |
Db    630 GGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAK 689

Qy    562 PAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT 621
          |  |      ||  |  |   :| | | ||  |   ||       |    :  || ::|| |
Db    690 PTTTT--TGAPGQPTTTTTGSPSKPTTTTTTKATTTT------TILNPIITTTTQKPTTT 741

Qy    622 TPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKE 681
          |      ||      |  |     ||  :|  ||    | ||   | |||      || |
Db    742 T------TT---KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKR 787

Qy    682 PAPTTPKKPAPK----ELAPTTTKEPTS--------------TTSDKPAPTTPKG-TAP 721
             ||   | |     |: |   ::                  ::|:| | :  |  |
Db    788 DEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIAD 847

Qy    722 TTPKEPA--------PTTPKEPAPTTPK------------------------------ 741
          |:   |         |  |    |  ||
Db    848 TSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDE 907

Qy    742 --GTAPTTLK----EPAPTTPKKPAPKELAPTTT--------------KGPTSTTSDKP 780
            |    ||      :|    | |   ||    :                |  | |
Db    908 TYGLPIDTLTGYPLDPVSLIPFNPETGELFDPISDEIMNGTIAGIVSGISASESLLSQKS 967

Qy    781 APTTPKE-----------TAPTTPKEPAPTTPKKPAPTTPETP-----PPTTSEVSTPTT 824
          ||  |                |    |    | :  | :||       ||  ||:
Db    968 APIDPATNMVVGEFGGLLNPATGVMIPGSLGPSEQTPFSPEIEDGGIIPP---EVAAANA 1024

Qy    825 TKEPTTIHKSPDESTPE----------------------LSAEPTPKALENSPKEPGVP 861
            |   :|  |  || ||                     :|  | | ::          |
Db   1025 DKFKLSIPPSVPESIPEKDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLN----P 1080

Qy    862 TTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA----TTTEKTTESKITA 917
          |||  |          |  | : :    : ||  |      |       |   | :|
Db   1081 IMKTPTQT---------DSVTGKPI----DPTTGLP-FNPPTGHLINPTNNNTMDSSFAG 1126

Qy    918 TTTQVTSTTTQDTTPFKITTLKTTTLAPK--VTTTKKTITTTEIMNKPEETAKPKDRAT- 974
                |   :    : :   : | | ||  |:       || |:::    | || :  |
Db   1127 AYKYAVSNGIKTDNVYGLPVDEITGL-PKDPVSDIPFNSTTGELVD--PSTGKPINNYTA 1183

Qy    975 ---------------NSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTS 1019
                          |     | : |      :  : :   |:       | | |    :
Db   1184 GIVSGKRGLPPIEDENGNLFDPSTKLPIDGNNQLVNPETNSTVSGSTSGSTKPKPGIPVN 1243

Qy   1020 ---TMPE---------------LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGG 1061
             :|:                : ||: | :  :   |   | | |     :||  |
Db   1244 GGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKV 1297

Qy   1062 AEGETPHMLLRPHVFMPEVTPD--------MDYLPRVPNQGIIINP-------------- 1099
            |  |  | |    |: | :         :  ||  |: | ||:|
Db   1298 IPGSLPGSLNYPSFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDE 1357

Qy   1100 MLSDETNICNGK---------------PVDGLTTLRNGTLV 1125
          :|::  ||   :                || ||  | |||||
Db   1358 ILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1398


RESULT   3
5202236-25
;Patent No. 5202236
;     APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;     TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;     NUMBER OF SEQUENCES: 39
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER: US/07/528,762
;       FILING DATE: 25-MAY-1990
;       APPLICATION NUMBER: 82,456
;       FILING DATE: 07-AUG-1987
;       APPLICATION NUMBER: 933,945
;       FILING DATE: 24-NOV-1986
;       APPLICATION NUMBER: 650,128
;       FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
;       LENGTH: 744
5202236-25


  Query Match             6.9%;  Score 506.5;  DB 6;  Length 744;
  Best Local Similarity   29.9%;  Pred. No. 5.2e-25;
  Matches  239;  Conservative   78;  Mismatches  308;  Indels  175;  Gaps   50;

Qy    212 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 271
          ||:|      |:|| ||                 | ::| |           | | |
Db     24 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 53

Qy    272 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 327
                      || |  |||:          |    ||:  | | |     | :  ||
Db     54 -----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    328 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 384
            | | || ||  ||   | | |: |: |    |  |  |    ||  | | |  |  |
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    385 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 436
            | |  |: |  |:  | |   ||  ||   | |    |  |    ||:  | | :|
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    437 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 485
             ||   :|:  ||   ||:  ||   :|:  ||   :|:  ||   :||  ||   :|:
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    486 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 538
            |:| |: |: |    |:| |: || |   ||   :|:  ||   :|    ||   ||
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    539 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 590
            :  ||   :|:  ||   :|:  ||   ||:  ||   :|:      :  ||   | |  :
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    591 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 639
            |  |: |    |  |: |    ||   :|:  ||   :|: | | ||    |:|    |
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    640 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 693
            |    ||   | |    |  |:    | : | |   | : |:| |  :   |   |
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    694 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 752
            ||   :||  :: |  ||      |: |    ||   :|:        ||   :|:
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    753 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 807
             ||  || |   ||    ||  ||   ||: :   :  |: |    ||   ||:  ||
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    808 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 854
              |  ||   :: | | | |     | |    |       :: |   |:|| |   :
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    855 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 900
            |  |  || |   :        ||    | | | | :   | | |  | |         | |
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719

Qy    901 KETATTTEKTTESKITATTT 920
            :   |   | ::| |  :|
Db    720 YKAKPTYPSTYKAKPTYPST 739


RESULT   4
US-08-700-651-5
```

```
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5

  Query Match             6.8%;  Score 499.5;  DB 3;  Length 1721;
  Best Local Similarity   21.8%;  Pred. No. 4e-24;
  Matches  300;  Conservative   94;  Mismatches  510;  Indels  469;  Gaps   44;

Qy      93 STTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKS------ 146
           | |    | |||: |                :| ::       |  ||| ::: |
Db      38 SITSGERKDPNKQAT--------------ISGSRSCGWKQGYSIDSSTGFRVDSITGLPT 83

Qy     147 --SKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQ 204
             |   |        |  |  :|:  |  |       ||    | : | | |  :|
Db      84 DPYSNCPFN-------PVTGNLVSRSTGKTIPN----TYAGVYRSN-ETKTTEPSANT-- 129

Qy     205 HNKVSTSPKITTAKP--------INPRPSLPPNSDTSKE---------------TSLTVN 241
           :  |  | :  | :|        ::|: : | ||: | |               |
Db     130 YAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQGQIFDMGSKVYIPYTKCVGV 189

Qy     242 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 301
           | ||  |  ||||   |:|                                      |||
Db     190 KHTTTTTTTTTTTTTTTT-------------------------------------TTT 211

Qy     302 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 361
             ||    | ||     :||    | ||  :  ||       ||| :  ||       |
Db     212 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 271

Qy     362 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEP 421
           ||     ||      |||     ||| :  || |:|  ||      ||      ||
Db     272 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTT 331

Qy     422 TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP 481
           | || |:|  ||     ||               ||      ||      |||   : ||
Db     332 TTTTTKKPTTTTTTTTTT-------------TTTTTTTTTTTTTTTTTTTTTTTTTT 377

Qy     482 KEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKE---------------------------- 513
            :|  |||  :  ||||:|  |||  :  ||  |
Db     378 TKPTTTTTTTTTTTTKKPTTTTTATTTTTSETESVIKPDEWCWLEKNGECEAKGATYVG 437

Qy     514 -----------------PSPTT------------------------------------TKE 521
                            |:  |                                     |
Db     438 VIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCRKGAGKLEFPDRSLDFTIP 497

Qy     522 P--------------------APTTPKEPAPTTPKKPAPTTPKE--------------- 545
           |                    :|   |: :  :    ||  | |
Db     498 PVAGHNSCSIIVGVSGDGKIHVSPYGSKDVSLIS----APIQPSELFNEVYCDTCTAKYG 553

Qy     546 ---PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKP 602
              |  :   ||| ||  |     ||  |  |   :| | | ||  |   ||
Db     554 AIHSGYQTSADFVTTTTAKPTTTT--TGAPGQPTTTTTGSPSKPTTTTTTKATTTT---- 607

Qy     603 APTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTT 662
             ||    :  || ::|| ||      ||      |  |     ||  :|  ||    | ||
Db     608 --TTLNPIITTTTQKPTTTT------TT---KVPGKPPIATTTTTLKPIVTTTTTKATTT 656

Qy     663 PKETAPTTPKGTAPTTLKEPAPTTPKKPAPK----ELAPTTTKEPTS------------ 705
           | |||       || |    ||   | |      |: |   ::
Db     657 TTTTVPTT-----TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGL 711

Qy     706 --TTSDKPAPTTPKG-TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKE---PAPTTPKKP 759
             ::|:| | :  |  | |:   |  |       |  |     |   |      | | :
Db     712 LLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTM 771

Qy     760 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEV 819
           :   ::    |  |  | :      | :|    | :|    |    | | |   | : |:
Db     772 SGLSVSYLAAKNLTVDTDETYG--LPIDTLTGYPLDPVSLIPFN--PETGELFDPISDEI 827

Qy     820 STPT--------TTKEPTTIHKS-------------------------------PDEST- 839
           |        :  |     ||                                   | | |
Db     828 MNGTIAGIVSGISASESLLSQKSALIDPATNMVVGEFGGLLNPATGVMIPGFLGPSEQTQ 887

Qy     840 ------------PELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTER--- 884
                       ||::|    |   : |  | ||       : ||     ||:  :
Db     888 FSPEIEDGGIIPPEVAAANADKFKLSIP--PSVP------ESIPE-----KDQKIDSISE 934

Qy     885 --------------DLRTTPETTTAAPKMTKETATTTEKTTESKITATT-TQVTSTTTQD 929
                          |  | :      :| | |:  |   |  ||        |
Db     935 LMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTDSVTGKPIDPTTGLPFNPPTGHL 994

Qy     930 TTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKP--EETAKPKD-------RATNSKATT 980
             |    |: ::           | |  :   |  | |  |||         :|  :
Db     995 INPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVGEITGLPKDPGSDIPFNSTTGELVD 1054

Qy     981 PKPQKPTKAPKKPTSTKKPKTMP--------------------RVRKPKTTPTPRKMTS 1019
           |   ||         : ||   |                      ::  |:|  |    ||
Db    1055 PSTGKPINNSTAGIVSGKPGLPPIEDENGNLFDPSTNLPIDGNNQLVNPETNSTVSGSTS 1114

Qy    1020 --TMPE--------------------------LNPTSRIAEAMLQTTTRPNQTPNSKL 1049
             | |:                         : ||: | :  :  |   | | |
Db    1115 GTTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNI-- 1172

Qy    1050 VEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD--------MDYLPRVPNQGIIINP-- 1099
             :||  |       |  |  |  |    |: | :        :  ||  |: | ||:|
Db    1173 --INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPAT 1228

Qy    1100 ------------MLSDETNICNGK---------------PVDGLTTLRNGTLV 1125
                      :|::  ||   :                 || ||  | |||||
Db    1229 KLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1281


RESULT   5
US-09-103-429A-3
; Sequence 3, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
```

```
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  786 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             6.8%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 1.8e-24;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy     293 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 352
Db      93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy     353 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 412
Db     136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy     413 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 469
Db     184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy     470 ----------------------------------------------PTTTKEPSPTTPK- 482
Db     242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy     483 ----EPAPTTT---------------------------------------------------- 489
Db     302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy     490 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 522
Db     362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy     523 APTT-----PKEP----------------------APTTPKKPAPTT----------- 542
Db     421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy     543 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 599
Db     481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy     600 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 657
Db     534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy     658 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 717
Db     588 AAPTAAPTAAPTTAVPEIPTTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy     718 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 777
Db     633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH----------APNTTAAPVTTTS 677

Qy     778 DKPAPTTPKETAPTTPKEPAPTTPKKP 804
```

```
Db     678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   6
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6


  Query Match             6.8%;  Score 498.5;  DB 3;  Length 1721;
  Best Local Similarity   21.8%;  Pred. No. 4.6e-24;
  Matches  300;  Conservative   94;  Mismatches  510;  Indels  469;  Gaps   44;

Qy      93 STTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSSTIWKIKS------ 146
Db      38 SITSGELKDPNKQAT-------------ISGSRSCGWKQGYSIDSSTGFRVDSITGLPT 83

Qy     147 --SKNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQ 204
Db      84 DPYSNCPFN-------PVTGNLVSRSTGKTIPN----TYAGVYRSN-ETKTTEPSANT-- 129

Qy     205 HNKVSTSPKITTAKP--------INPRPSLPPNSDTSKE--------------TSLTVN 241
Db     130 YAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQGQIFDMGSKVYIPYTKCVGV 189

Qy     242 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 301
Db     190 KHTTTTTTTTTTTTTTTTT----------------------------------------TTT 211

Qy     302 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 361
Db     212 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 271
```

```
Qy    362 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEP 421
Db    272 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTT 331

Qy    422 TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP 481
Db    332 TTTTTKKPTTTTTTTTTTT-------------TTTTTTTTTTTTTTTTTTTTTTTTTTT 377

Qy    482 KEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKE---------------------------- 513
Db    378 TKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVG 437

Qy    514 -----------------PSPTT-----------------------------------TKE 521
Db    438 VIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCRKGAGKLEFPDRSLDFTIP 497

Qy    522 P--------------------APTTPKEPAPTTPKKPAPTTPKE--------------- 545
Db    498 PVAGHNSCSIIVGVSGDGKIHVSPYGSKDVSLIS----APIQPSELFNEVYCDTCTAKYG 553

Qy    546 ---PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKP 602
Db    554 AIHSGYQTSADFVTTTTAKPTTTT--TGAPGQPTTTTTGSPSKPTTTTTTKATTTT---- 607

Qy    603 APTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTT 662
Db    608 --TTLNPIITTTTQKPTTTT------TT---KVPGKPPIATTTTTLKPIVTTTTTKATTT 656

Qy    663 PKETAPTTPKGTAPTTLKEPAPTTPKKPAPK----ELAPTTTKEPTS------------- 705
Db    657 TTTTVPTT-----TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGL 711

Qy    706 --TTSDKPAPTTPKG-TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKE---PAPTTPKKP 759
Db    712 LLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTM 771

Qy    760 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEV 819
Db    772 SGLSVSYLAAKNLTVDTDETYG--LPIDTLTGYPLDPVSLIPFN--PETGELFDPISDEI 827

Qy    820 STPT--------TTKEPTTIHKS------------------------------PDEST- 839
Db    828 MNGTIAGIVSGISASESLLSQKSALIDPATNMVVGEFGGLLNPATGVMIPGFLGPSEQTQ 887

Qy    840 ------------PELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTER--- 884
Db    888 FSPEIEDGGIIPPEVAAANADKFKLSIP--PSVP------ESIPE-----KDQKIDSISE 934

Qy    885 --------------DLRTTPETTTAAPKMTKETATTTEKTTESKITATT-TQVTSTTTQD 929
Db    935 LMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTDSVTGKPIDPTTGLPFNPPTGHL 994

Qy    930 TTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKP--EETAKPKD-------RATNSKATT 980
Db    995 INPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVGEITGLPKDPGSDIPFNSTTGELVD 1054

Qy    981 PKPQKPTKAPKKPTSTKKPKTMP--------------------RVRKPKTTPTPRKMTS 1019
Db   1055 PSTGKPINNSTAGIVSGKPGLPPIEDENGNLFDPSTNLPIDGNNQLVNPETNSTVSGSTS 1114

Qy   1020 --TMPE---------------------------LNPTSRIAEAMLQTTTRPNQTPNSKL 1049
Db   1115 GTTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNI-- 1172

Qy   1050 VEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD--------MDYLPRVPNQGIIINP-- 1099
Db   1173 --INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPAT 1228

Qy   1100 -----------MLSDETNICNGK---------------PVDGLTTLRNGTLV 1125
Db   1229 KLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1281
```

```
RESULT   7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/07/638,431
;      FILING DATE:  19910110
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avrom D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-07-638-431-2
```

Query Match 6.7%; Score 488.5; DB 1; Length 826;
Best Local Similarity 26.4%; Pred. No. 8.5e-24;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    329 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 381
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNN 332

Qy    382 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 437
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    438 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 495
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    496 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 553
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    554 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 612
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    613 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 671
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564
```

```
Qy    672 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 731
                          |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    732 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 791
          ||||                |  : ||:|    : |   : |     :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    792 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 845
            | |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    846 PTPKALENSPK-------------------EPGVPTTK 864
            |     |  ||                  || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene encoding
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  PCT/US92/00018
;      FILING DATE:  19920103
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avram D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
PCT-US92-00018-2


  Query Match             6.7%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 8.5e-24;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    329 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 381
          ||  ::  |  |  |  |      |: |:||  |    :|     |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNN 332

Qy    382 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 437
          |      : |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |:|:     |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    438 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 495
```

```
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    496 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 553
          :| |    : |: | :|:|          |  | ||: |  |  | ||: | |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    554 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 612
          |:        |  |  |  |: | |  |: |||                  |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    613 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 671
          : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    672 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 731
                          |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    732 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 791
          ||||                |  : ||:|    : |   : |     :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    792 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 845
            | |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    846 PTPKALENSPK-------------------EPGVPTTK 864
            |     |  ||                  || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   9
5202236-13
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;      LENGTH: 652
5202236-13


  Query Match             6.5%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 3.8e-23;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy    212 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 271
          ||:|      |:|| ||                  | ::| |             | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    272 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 327
                  || |  |||:        |    ||:  | | |      | :  ||
Db     41 -----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy    328 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 384
          | | || ||  ||   | | |: |:  |    |  |  |    ||   | | |  |  |
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    385 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 439
          | |  |: |  |:  | |   ||  ||   | |    | |    | :  | | :|  |
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196
```

```
Qy    440 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 490
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    491 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 537
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308

Qy    538 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 593
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    594 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 648
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    649 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 701
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    702 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 759
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    760 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 812
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    813 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 872
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    873 MTTTAKDKTTERDLRTTPETTTAAP 897
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645
```

RESULT 10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
; APPLICANT: VAN HEERDE, GEORGE V.
; APPLICANT: VAN RIJN, ALEXIS C.
; APPLICANT: BOUWSTRA, JAN B.
; APPLICANT: DE WOLF, FREDERIK A.
; APPLICANT: MOOBROEK, ANDREAS
; APPLICANT: WERTEN, MARC W.T.
; APPLICANT: WIND, RICHELE D.
; APPLICANT: VAN DEN BOSCH, TANJA J.
; TITLE OF INVENTION: SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
; TITLE OF INVENTION: SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
; TITLE OF INVENTION: PREPARATION THEREOF
; FILE REFERENCE: 2728-2
; CURRENT APPLICATION NUMBER: US/09/219,849
; CURRENT FILING DATE: 1998-12-23
; NUMBER OF SEQ ID NOS: 50
; SOFTWARE: PatentIn Ver. 2.1
; SEQ ID NO 5
; LENGTH: 960
; TYPE: PRT
; ORGANISM: Artificial Sequence
; FEATURE:
; OTHER INFORMATION: Description of Artificial Sequence: Illustrative
; OTHER INFORMATION: amino acid sequence
US-09-219-849-5

Query Match 6.2%; Score 452; DB 4; Length 960;
Best Local Similarity 25.6%; Pred. No. 2.2e-21;
Matches 173; Conservative 60; Mismatches 309; Indels 134; Gaps 41;

```
Qy    293 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 350
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    351 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 406
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    407 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA---------------- 445
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    446 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 491
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    492 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 548
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    549 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 598
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    599 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 653
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    654 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 696
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    697 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 751
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    752 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 803
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    804 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 854
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    855 PKEPGVPTTKTPAATK 870
Db    696 SRDPGPPGAHGPAGPK 711
```

RESULT 11
US-09-041-886-23
; Sequence 23, Application US/09041886
; Patent No. 6235872
; GENERAL INFORMATION:
; APPLICANT: Bredesen, Dale E.
; APPLICANT: Rabizadeh, Sharroz
; TITLE OF INVENTION: Proapoptotic Peptides, Dependence
; TITLE OF INVENTION: Polypeptides and Methods of Use
; NUMBER OF SEQUENCES: 72
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Campbell & Flores LLP
; STREET: 4370 La Jolla Village Drive, Suite 700
; CITY: San Diego
; STATE: California
; COUNTRY: United States
; ZIP: 92122
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.25
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/09/041,886

```
;      FILING DATE:
;      CLASSIFICATION:
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Campbell, Cathryn A.
;      REGISTRATION NUMBER:  31,815
;      REFERENCE/DOCKET NUMBER:  P-LJ 2626
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (619) 535-9001
;      TELEFAX:  (619) 535-8949
;  INFORMATION FOR SEQ ID NO:  23:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1185 amino acids
;      TYPE:  amino acid
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-09-041-886-23


  Query Match             5.8%;  Score 425.5;  DB 4;  Length 1185;
  Best Local Similarity   21.8%;  Pred. No. 1.5e-19;
  Matches  240;  Conservative  119;  Mismatches  437;  Indels  303;  Gaps   52;

Qy      88 SQTIKSTTKRSPKPPNKKKTKKVIESEEITEEHSVSENQESSSSSSSSSSSTIWKIKSS 147
           | :::|  |:    |           ||:      |    |:|||          | : |
Db       9 SMSMRSGRKKEAPGPR----------EELRSRGRASPGGVSTSSSDG--------KAEKS 50

Qy     148 KNSAANRELQKKLKVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK 207
           : :|              || | ::  |||   |:: |             :  :   ::
Db      51 RQTA-------------KKARVEEASTPK---VNKQGRS----------EEISESESE 82

Qy     208 VSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTT 267
            : :|| |   :   |||  | : |:    ||  | : : : ::    |: ||
Db      83 ETNAPKKTKTEQELPRPQSPSDLDSLDGRSL--NDDGSSDPRDIDQDNRSTS-------P 133

Qy     268 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTP--KEPASTTP 325
           |    |:|  |                 :   ::|||   |  | |  | .:|  :||
Db     134 SIYSPGSVENDS--------------DSSSGLSQGPA--RPYHPPPLFPPSPQPPDSTP 176

Qy     326 KEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPT 385
           ::|  :          || |: | :    | || ||:    ||  |  | |      |  |
Db     177 RQPEASF---------EPHPSVTPTGYHAPMEP-PTSRMFQAP--PGAPPPHPQLYPGGT 224

Qy     386 --TTKSAPTTPK---------------EPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPK 427
             |  ||                   | ||||   : :      || |    ||||
Db     225 GGVLSGPPMGPKGGGAASSVGGPNGGKQHPPPTTPISVSSSGASGAPPTKP----PTTPV 280

Qy     428 EPAPTTKEPAP-----TTPKEPAPTA--PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTT 480
               | |       ||  | | |  |   |  :|    |    :| |     |||
Db     281 GGGNLPSAPPPANFPHVTPNLPPPPALRPLNNASASP----PGLGAQPLPGHL--PSPYA 334

Qy     481 PKE------PAPTTTKSAPTTTKEP--APTTTKSAPTTP-KEPSPTTTKEPAPTTPKEPA 531
             :       | |   :  ||    |   |  : |||  | : |  :::   |  :    :
Db     335 MGQGMGGLPPGP---EKGPTLAPSPHSLPPASSSAPAPPMRFPYSSSSSSSAAASSSSSS 391

Qy     532 PTTPKKPAPTTPKEPAPTTPKE-PAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLT--- 587
            ::   | | :  :  |: |   | ||:      :   :|    || | |: | :
Db     392 SSSSASPFPAS--QALPSYPHSFPPPTSL-----SVSNQP----PKYTQPSLPSQAVWSQ 440

Qy     588 --PTTPE--KLAPTTPEKPAPTTPEELAPTTPEEPTPT---------------------- 621
             | |   :|  :   | |  |    | :|   |  |
Db     441 GPPPPPPYGRLLANSNAHPGPFPPSTGAQSTAHPPVSTHHHHHQQQQQQQQQQQQQQHHG 500

Qy     622 --TPEEPAP-----------TTPKAAAPN----TPKEPAPTTPKEP------APTTPKE 657
             |  |                | | :|:      |  | |      |       :   |
Db     501 NSGPPPPGAFPHPLEGGSSHHAHPYAMSPSLGSLRPYPPGPAHLPPPHSQVSYSQAGPNG 560

Qy     658 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP---TTTKEPTSTTSDKPAPT 714
           |  ::   :: :| :|: | :   |:|:   : ||    |    ||   | :|      :
Db     561 PPVSSSSNSSSSTSQGSYPCS--HPSPSQGPQGAPYPPPPVPTVTTSSATLSTVIATVAS 618

Qy     715 TPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTS 774
           :| |    :|  | |   : |:|   |   |   :|     |: :   |  :|
Db     619 SPAGYKTASPPGPPPYGKRAPSPGAYKTATPPGYKPGSP-----PSFRTGTPPGYRG--- 670

Qy     775 TTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKS 834
             ||    | | |  :||    | |  |  |:   |  |||  :  | | :   | | :
Db     671 -TSPPAGPGTFKPGSPTV--GPGPLPPAGPS-GLPSLPPPPAAPASGPPLS--ATQIKQE 724

Qy     835 PDE--STPELSAEPTPKALENS--PKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP 890
           | |   |||    | | |   |  ||   ||:  : :|   :       :     ||   |
Db     725 PAEEYETPE---SPVPPARSPSPPPKVVDVPSHASQSARFNKHLDRGFNSCARSDLYFVP 781

Qy     891 ETTTAAPKMTKETATTTEKT-TESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTT 949
                 |: |: |   ||   |::  |
Db     782 ---LEGSKLAKKRADLVEKVRREAEQRAR------------------------------ 807

Qy     950 TKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK 1009
                       ||  : ::|    :    | ::  :: |    : |     | |
Db     808 -------------EEKERERERERERERERERELERSVKLAQEGRAP-----VECPS 848

Qy    1010 TTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM 1069
           | | :     |   | | :|      |  |   |::  :     || |
Db     849 LGPVPHR-----PPFEPGSAVA------TVPPYLGPDTPALRT--LSEYA---------- 885

Qy    1070 LLRPHVFMPEVTPDMDYLP 1088
             ||||  |       |:|
Db     886 --RPHVMSPGNRNHPFYVP 902


RESULT  12
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
```

```
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             5.8%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 2.9e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    236 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 294
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    295 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 337
          |    :||:|    |    ||               | |    ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    338 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 378
             :||| :|         |||     |             ||      |  | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    379 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 425
            :  | | :|        |||     |  | |      ||      |  | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    426 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 472
             :  | | :  |        ||      |  |||        ||      |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    473 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 521
            :  |  |   ||      |||     |  |        |||     |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    522 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 573
            ||      |  | |       ||      |  | |       ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    574 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 629
          :  :  |    | |      |:    ||      |  | |      |: :  |     |  |

Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    630 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 684
            | |     :||:     |  | |      |:  :  |  |  |||      | ||     |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    685 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 732
            | |              :| |    ||      ||    :| |   | |      |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    733 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 778
              |  | |      :||      |  | |           :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    779 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 822
            :| |  | |       :||      |  | |           :| | :  | |       | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    823 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 868
              | :        :  :||       |  |     : ::|      | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    869 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 924
             :|   :||       |:  | |  | :|  ||      |:    :                |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    925 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 984
                |       |   |  ||      :  :    |  ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    985 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1044
             : ||      |  |   |       | |    :||  |:  |               |  |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1045 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1103
           | :  |   |      |:     |            ||   |  |   :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1104 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1158
            |:  : :|  | |    :|   |            | | |          |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  13
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
```

```
;       APPLICATION NUMBER:  08/479,537
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5


  Query Match             5.8%;  Score 424.5;  DB 4;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 2.9e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy     236 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 294
           |  |||           |  :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     295 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 337
           |    :||:|    |    ||                | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     338 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 378
             :||| :|         |||      |             ||     |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     379 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 425
             :  | | :|         |||      |  | |        ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     426 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 472
             :  | | :  |         ||      |  |||        ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     473 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 521
             :  |  |   ||       |||      |  |         |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     522 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 573
           ||      |  | |         ||      |  | |        ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     574 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 629
             :  :  |    | |       |:     ||      |  | |      |:  :  |     |  |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     630 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 684
             | |     :||:    |  | |     |:  :  |  |  |||      | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     685 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 732
           | |                :| |   ||       ||     :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     733 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 778
             |  | |       :||      |  | |           :| |    ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     779 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 822
           :| |  | |       :||      |  | |           :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     823 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 868
             | :       :  :||      |  |     : ::|      | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     869 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 924
            :|   :||       |:  | |  | :|  ||      |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     925 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 984
               |       |  |  ||       :  :    |  ||      |::     | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     985 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1044
             : ||      |  |   |       | |    :||  |:  |                  | |
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    1045 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1103
            | :  |   |       |:     |                ||   |  |   :      |
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy    1104 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1158
            |:  : :|  | |   :|  |                |  | |         |: |  |
Db     958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  14
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;     APPLICANT:  CHAMBON, Pierre
;     APPLICANT:  KIENY, Marie-Paule
;     APPLICANT:  LATHE, Richard
;     APPLICANT:  HAREUVENI, Mara
;     TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;     TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
```

```
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
```

US-08-479-537A-2

Query Match 5.8%; Score 424.5; DB 2; Length 2035;
Best Local Similarity 23.7%; Pred. No. 3.2e-19;
Matches 255; Conservative 69; Mismatches 511; Indels 241; Gaps 53;

```
Qy    236 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 294
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    295 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 337
          |    :||:|    |    ||                | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    338 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 378
          :||| :|         |||     |                ||     | | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    379 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 425
          :  | | :|         |||     | | |        ||     | | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    426 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 472
            : | | :  |         ||      | |||        ||     | | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    473 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 521
          :  |  |   ||       |||     | |                |||     | |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    522 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 573
          ||      | | |        ||      | | |        ||      | |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    574 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 629
          :  :  |    | |      |:    ||      | | |      |:  :  |     | |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    630 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 684
          | |    :||:      |  | |      |:  :  | |  |||      | ||      |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    685 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 732
          | |           :| |   ||          ||     :| |  | |     |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    733 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 778
          |  | |      :||      |  | |                :| |   ||    | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    779 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 822
          :| |  | |       :||      |  | |              :| | :  | |     | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    823 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 868
            | :       :  :||      |  |    : ::|     | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    869 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 924
           :|   :||        |:  | | |  | :|  ||     |:    :              |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    925 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 984
                |       |   |  ||      :   :      |  ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    985 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1044
            : ||      |  |   |       | |    :||  |:  |                | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1045 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1103
           | :  |   |      |:    |              ||   | |  :      |
```

```
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1104 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1158
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  15
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2


  Query Match             5.8%;  Score 424.5;  DB 4;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 3.2e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    236 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 294
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    295 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 337
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    338 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 378
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    379 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 425
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    426 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 472
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    473 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 521
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    522 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 573
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    574 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 629
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    630 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 684
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    685 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 732
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    733 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 778
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    779 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 822
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702
```

```
Qy    823 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 868
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    869 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 924
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    925 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 984
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    985 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1044
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX----------PGST 911

Qy   1045 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1103
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1104 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1158
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

Search completed: April 26, 2002, 16:26:07
Job time: 507 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:32:20 ; Search time 114.61 Seconds
(without alignments)
904.577 Million cell updates/sec

Title: AA7
Perfect score: 7323
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1361

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

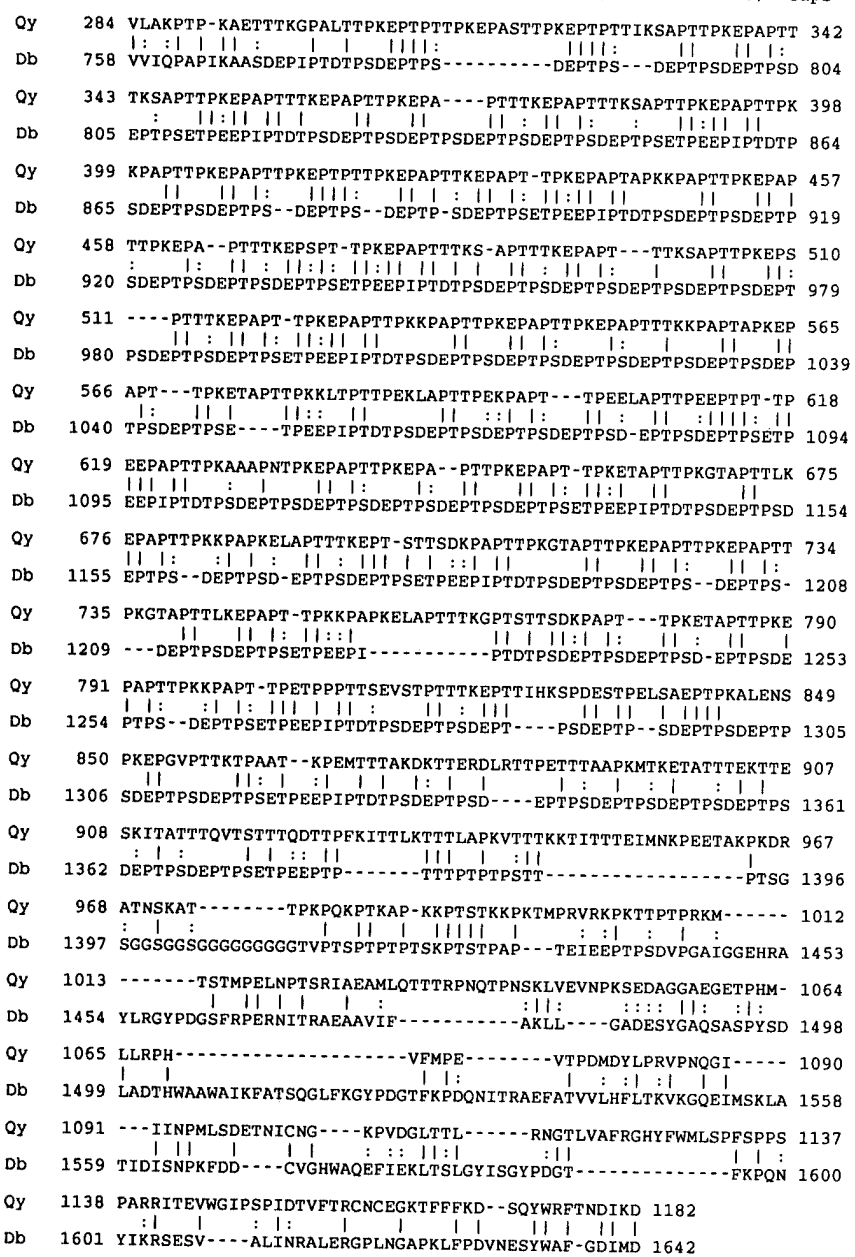
Result No.	Score	% Query Match	Length	DB	ID	Description
1	1088.5	14.9	3020	2	A43932	mucin 2 precursor,
2	950	13.0	1664	2	T18262	S-layer protein -
3	867	11.8	1489	2	T31108	cyst germination s
4	835	11.4	1274	2	T16251	hypothetical prote
5	828.5	11.3	2187	2	T30826	nascent polypeptid
6	800	10.9	1188	2	S49915	extensin-like prot
7	795.5	10.9	1367	1	S48478	glucan 1,4-alpha-g
8	704.5	9.6	1151	2	T18535	high molecular mas
9	674	9.2	1229	2	T25697	hypothetical prote
10	672	9.2	1344	1	A35175	mucin 1 precursor,
11	670	9.1	7962	2	I38346	elastic titin - hu
12	668	9.1	3507	2	T34513	hypothetical prote
13	663	9.1	3570	2	T45025	mucin MUC5B, trach
14	646	8.8	990	2	I51618	nucleolar phosphop
15	633	8.6	489	2	T11622	extensin class 1 p
16	632	8.6	761	2	C84672	hypothetical prote
17	632	8.6	971	2	T19431	hypothetical prote
18	626	8.5	6642	2	T29757	protein UNC-89 - C
19	622.5	8.5	839	2	F75518	hypothetical prote
20	607.5	8.3	801	2	T29018	hypothetical prote
21	607.5	8.3	924	2	S27923	gene LF3 protein -
22	605	8.3	379	2	S50125	larval glue protei
23	568	7.8	2232	2	T34434	hypothetical prote
24	566.5	7.7	875	2	S23760	polyphenolic adhes
25	565	7.7	1630	2	A53577	ascites sialoglyco
26	559.5	7.6	350	2	S22456	hydroxyproline-ric
27	556.5	7.6	856	2	T16543	hypothetical prote
28	551	7.5	620	2	S06733	hydroxyproline-ric
29	545	7.4	873	2	A47283	calphotin - fruit
30	542	7.4	369	2	S20500	hydroxyproline-ric
31	538.5	7.4	416	2	JU0465	extensin precursor
32	535	7.3	756	2	T27642	hypothetical prote
33	533	7.3	1087	1	QFMSH	neurofilament trip
34	530.5	7.2	1162	2	JH0557	exo-alpha-sialidas
35	530	7.2	865	2	A47282	calcium-binding pr
36	528	7.2	3534	2	T42567	tegument protein 2
37	527.5	7.2	1459	2	T32271	hypothetical prote
38	522	7.1	328	2	JQ0985	hydroxyproline-ric
39	519	7.1	813	2	S70795	vsaA protein precu
40	518	7.1	866	2	T45462	membrane glycoprot
41	518	7.1	1072	1	A37221	neurofilament trip
42	514	7.0	1611	2	T38236	hypothetical prote
43	512.5	7.0	867	2	T45463	membrane glycoprot
44	510.5	7.0	1832	2	T31113	mucin-like glycopr
45	509.5	7.0	662	2	A45155	mucin FIM-C.1 - Af

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

```
J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

  Query Match             14.9%;  Score 1088.5;  DB 2;  Length 3020;
  Best Local Similarity   28.2%;  Pred. No. 2.7e-38;
  Matches  379;  Conservative   98;  Mismatches  514;  Indels  355;  Gaps   51;

Qy      43 TCNCDY--NCQHYMECCPDFKRVCTAELSCKGRCFESFERGRECDCDAQCKKYDKCCPDY 100
           :|:||   :|: :     : : || | :|       | |  :  |   |  |:   | :
Db    1079 SCSCDTGGDCECFCSAVASYAQECTKEGAC------VFWRTPDL-CPIFCDYYN---PPH 1128

Qy     101 ESFCAEVHNPTSPPSSKKAPPPSGASQTI-----KSTTKRSPK--PPNKKKTKKVIESEE 153
           |  |   : |    | :    :|    |      :  | ||  |  ::  || : :::
Db    1129 E--CEWHYEPCGNRSFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADK 1186

Qy     154 I-TEVKDNKKNRTKKKPTPK----------PPVV--DEAGSGLD---NGDF---KVTTPD 194
             |:|         || :             ||   | |  |:   :| |   ::  |:
Db    1187 CGCYVEDTHYPPGASVPTEETCKSCVCTNSSQVVCRPEEGKILNQTQDGAFCYWEICGPN 1246

Qy     195 TSTTQH---NKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKET--TTT 249
           :  :|     ::| |   |       |: |         || |    ||  |   |  :||
Db    1247 GTVEKHFNICSITTRPSTLTTFTTITLPTTP--------TSFTTTTTTTTPTSSTVLSTT 1298

Qy     250 NK--------------QTSTDGKEKTTS--AKETQSIEKTSAKD--------------- 277
           |               | ||  :         : ||  | ||
Db    1299 PKLCCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCD 1358

Qy     278 ----------------------------------LAPTSKVLAKPTPKAETTTKGPA 300
                                               :   |     |:|   |||  |
Db    1359 VSVGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPP 1418

Qy     301 LTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-TTPKEPAPTTTKSAPTTPKEPAPTTTK 359
           |||  || ||   | :|||  |  ||     | |||  |  |||   |||   | ||||
Db    1419 TTTPSPPTTTTTTPPPTTTPSPPITTTTTPLPTTTPSPPISTTTTPPPTTTPSP-PTTTP 1477

Qy     360 EPAPTTPKEPAPTTTKEPAPTTTKSAP-TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTP 418
           |  |||  |  |||  | ||||  | |  |||  |  :|   |  |||  |  ||   |
Db    1478 SPPTTTPSPPT-TTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTTTTPPPT 1536

Qy     419 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 478
           |||  |  |   | ||:     ||    | |||   | |||   |  |||  |   |
Db    1537 TTPSPPTTTPITP-PTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTPSPPTITTTT 1595

Qy     479 PAPTTTKSAP-TTTKEPAPTTTKSAP-TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPT 536
           | |||| | | |||  | ||||  | |  |||  |  :||  |  |||  |  ||   |  |
Db    1596 PPPTTTPSPPTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTT 1655

Qy     537 TPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK 596
           ||  |  |||  |  |||  |  | |  |  |||  : |||    | |||     |||
Db    1656 TPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTP--SPPTT----TMTTPS--PTTTPSS 1707

Qy     597 PAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPT 656
           |  ||    : |||  | |||   |:|||  :    |     |||   |  |||  |: |
Db    1708 PITTTTTPSSTTTP-SPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSIT 1766

Qy     657 TP-----KETAPTTP-----------KG---------------------------- 668
           |       | ||||             |
Db    1767 PPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRAT 1826

Qy     669 ------------------------------------------------------TAPTTLK 675
                                                                   | |||:
Db    1827 MYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTM- 1885

Qy     676 EPAPTTPKKPAPKELAPTT--TKEPTSTTSDKP----APT-TPKGTAPTTPKEPAPTTPK 728
             || : | |  :  ||  |  || |::  |    ||| ||  |  |    | ||  :
Db    1886 --TTTTTENPTPTPITTTTTVTPTPTPTSTQSPNGLQAPTPTPISTTTTVTPTPTPTGTQ 1943

Qy     729 EPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKPAPTTPKETAPTT 787
           |  |||  |  |    | ||  : |    :  |||  || | || |    ||  |  |
Db    1944 TPT-TTPITTTTTVTPTPTPTGTQTPTTVLITTTTTMTPTPTPTSTKSTTVTPITTTTTV 2002

Qy     788 PKEPAPTTPKKP-------APTTPETPPPTTSEVSTPTTTK-----EPTTIHKSPDESTP 835
             | ||  : |          |  || |||     || |      | || :  |: |||
Db    2003 TATPTPTGTQTPTMIPISTTTTVTPTPTPTTGSTGPPTHTSTAPIAELTTSNPPPESSTP 2062

Qy     836 ELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMT 895
           : |   :    |::     :|     :| | |  ||   |:     : | :|| :|  |
Db    2063 QTSRSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPTTTSGGHTLSPPPSTTTSPPGT 2122

Qy     896 KETATTTEKTTESKITATTTQVTSTTTQDTTPFKITT---LKTTTLAPKVTTTKKTITTT 952
             |||   : |  | :| | |:|:    ||  ::|   ::|| | |  ::
Db    2123 PTRGTTT--GSSSAPTPSTVQTTTTSAWTPTPTPLSTPSIIRTTGLRPYPSSVLICCVLN 2180

Qy     953 EIMNKPEE----------------------------TAKPKDRATNSKATTPKPQKPTK 983
           :    | |                             : |    | || :|| | ||:
Db    2181 DTYYAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTPSPTPTPSK-STPTPSKPSS 2239

Qy     984 APKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSK 1043
             | |||   ||   |    |:             : | |  : :  :  | : | |
Db    2240 TPSKPTPGTKPPECPDFDPPR------------QENETWWLCDCFM-ATCKYNNTVEIV 2285

Qy    1044 LVEVNP------------KSEDAGG-----------AEGETPHMLL----------R 1067
             ||  |              : ||  |              |: || :
Db    2286 KVECEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGD-PHYVTFDGLYYSYQGNC 2344

Qy    1068 PHVFMPEVTPDMDYLPRVPNQGIIIN 1093
             :| : |::| :|      | |: |:
Db    2345 TYVLVEEISPSVD------NFGVYID 2364


RESULT   2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

  Query Match             13.0%;  Score 950;  DB 2;  Length 1664;
  Best Local Similarity   31.0%;  Pred. No. 9.5e-33;
```

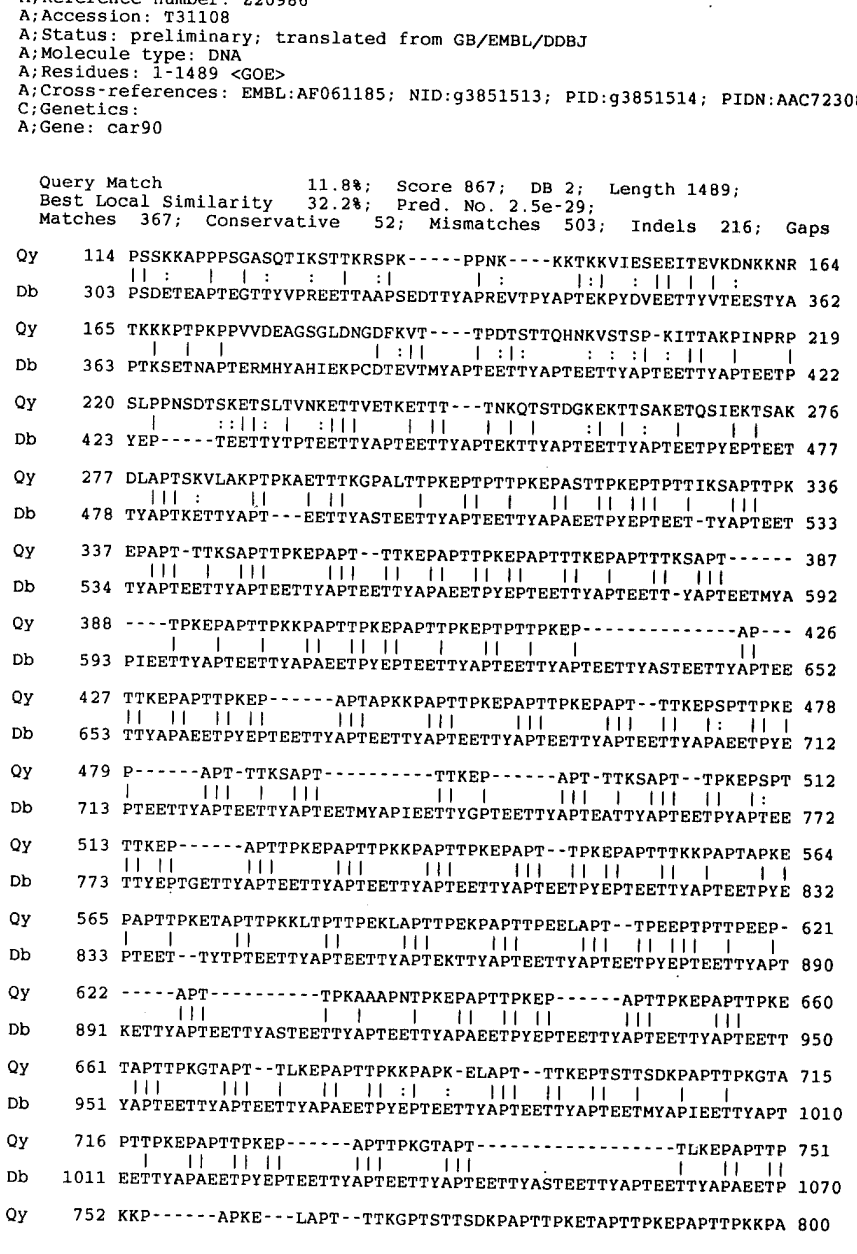
Matches 312; Conservative 104; Mismatches 361; Indels 230; Gaps 55;



RESULT 3
T31108

cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90

Query Match 11.8%; Score 867; DB 2; Length 1489;
Best Local Similarity 32.2%; Pred. No. 2.5e-29;
Matches 367; Conservative 52; Mismatches 503; Indels 216; Gaps 51;



```
Db   1071 YEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPT 1130

Qy    801 PTTPETPPPTTSEVSTPTTTKEPT--TIHKSPDESTPELSAEPTPKA-LENSPKEPG--- 854

Db   1131 EETTYAPTEETMYAPIEETTYGPTEETTY-APTEATTYAPTEETPYAPTEETTYEPTGET 1189

Qy    855 --VPTTKTPAATKPEMTTTAKDKTTERDLRTTP----ETTTAAPKMTKETATTTEKTTES 908

Db   1190 TYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPTEETTYAP-----TEETTYEPTEE 1244

Qy    909 KITATTTQVTSTTTQDTT----------PFKITT---LKTTTLAPKVTTTKKTITTTEIM 955

Db   1245 TTYAPTEETTYAPTEETTYAPTEETMYAPIDETTYGPTEETTYAPTEATTYAPTEETPYA 1304

Qy    956 NKPEETAKPKDRAT--NSKATTPKP-QKPTKAPKKPT-----------STKKP-KTMPR 999

Db   1305 PTEETTYEPTGETTYAPTEETTYAPTEETTYAPMEETPYEPAEESTSTVSTEKPCNTEEF 1364

Qy   1000 VRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP--NQTPNSKLVEVNPKSEDAG 1055

Db   1365 TDEPTDEPTDE--PSDEPTDEPTDEPTDLPTDEPSTPCDNQGINGIGVENKVRYNNAG 1420
```

RESULT 4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN0
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

Query Match 11.4%; Score 835; DB 2; Length 1274;
Best Local Similarity 27.5%; Pred. No. 4.7e-28;
Matches 310; Conservative 111; Mismatches 446; Indels 260; Gaps 60;

```
Qy      95 KCCPDYESFCAEVHNPTSPPSSKKAP-------------------PPSGASQTIKSTTKR 135

Db     189 KAAPSKE------HDPIVPPTPIKNPAKKWKPPWEDDEVPTEEIKEPEPATRKVPALKKK 242

Qy     136 SP----KP-PNKKKTKKVIESEEITEVKDNKKNRTKK-KP-----TP-----KPPVVDEA 179

Db     243 EPSTSVKPVSDPSPTKKVPVKKEPEVPPTPIKNPTKKWKPPWEDETPVEEVKEPPVPEKK 302

Qy     180 GSGLDNGD----FKVTTPDTSTTQHNKVSTSPKITTAKPI-NPRPSLPPNSDTSKETSLT 234

Db     303 APVLKKKDPAPAAKARDPSPSKAAPKKVEPSSPVVPPTPVKNPVKKYKPPWEVDDEPAEE 362

Qy     235 VNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAE- 293

Db     363 VKKPSAPEKKTPVLKRKEPEPSSTTPSSDPSPKKAAPAVKPRDSSPKKATPLQADPKAQE 422

Qy     294 ---TTTKGPA----------------LTTPKEPTPTTP----KEPASTTPKEP----TP 325

Db     423 VPPTPVKNPVKKYKPPWEVDDEDPVEEVKQPEAPAKKTPVLKRKEPAAKDTAKPATSKTP 482

Qy     326 -TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPA------------------PTT 365

Db     483 ETPEKKDPVKPRDSSPKKVAAKPDSAQAPA-TPVKNPVKKWRPPWEDDETPADDVSKPTD 541

Qy     366 PKEPAPTTTKEPAPT-------TTKSAPTTPKEP-----APTTPKKPAPTTPKE----PA 409

Db     542 AKKTPSLAKKDPAPAKESLKPKADTKAPAKPRDPSPKKVAPTAPEKKTPVLAKKEPAGPA 601

Qy     410 PTTPKEPTPTTPKEPAPTTKEPAPTTPK-EPAPTAPKKPAP------TTPKEPAPTTPKE 462

Db     602 DSKTKEPEKSKPRDPSPKKAVPAKPVPKTEVAPAAVKKPEPISKPKDTAPKKAEPNSPVV 661

Qy     463 PAPTTTKEP----SPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTP-KEPSPTTTKEP 517

Db     662 P-PTPVKNPVKKWKPPWEDDDAPAKPVSLPEPEKK-TPVLAKKAPTKPDSEAAADPVSGP 719

Qy     518 APTTPK--EPAPTTPKKPAP--TTPKEPAPTTPKEPAPTTTKKPAPTA-PKEPAPTTPKE 572

Db     720 SSKDPKLAKKAPVKPRDPSPMKAVPIKPAPKT--EVPPAVVKKPEPVAKSRDPSPKKAK- 776

Qy     573 TAPTTPKKLTPTTP-----EKLAPTTPEKPAPTTPEELAPTTPEEP-----TPTTPEEPA 622

Db     777 AEPNSP--VVPPTPVKNPVKKWKPPWEDDDAPAEPVNVPEPEKKTPVLAKKTPVKPRDPS 834

Qy     623 PTTPKAAAPNTPKEPAPTTPKEPAPTT-PKEPAPTTPKETAPTTPKGTAPTTLKEPAPTT 681

Db     835 PKKAVPAKPSTKTDAPPVSVKKPEPVSKPKEPSPKKAEPNSPVVP----PTPVKNPVKKW 890

Qy     682 PKKPAPKELAPT-TTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAP 740

Db     891 -KPPWEDDDEPTEEVKKPSE--PEKKTPVLAK-KEPEKPKD-APKVAAKPRDPSPKKAVP 945

Qy     741 TTLKEPAPT-------TPKKPAP-----KELAPTTTKGPTS------TTSDKP-----AP 777

Db     946 E--KEPAKVAAKPRDLSPKKAIPIPANTQEAPPTPVKNPVKKWKPPWEDDDEPAEPVSAP 1003

Qy     778 TTPKET------APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSP- 830

Db    1004 EPEKKTPVLAKKAPAKPRDP---SPKKAAPVAAK-PDPKIPEV-PPTPVKNPVKKWKPPW 1058

Qy     831 ---DESTPELSA-EP---TPKALENSPKEPGV-------------PTTKTPAAT------ 864

Db    1059 EDDDEPSEPVSAPEPEKKTPVLAKKAPTKPATKPDSEAAADPVSGPTSKDPKLSKKAPVE 1118

Qy     865 KPEMTTTAKDKTTERDLRTTPETT--TAAPKMTKETATTTEKTTESKITATTTQVTSTTT 922

Db    1119 KPKPTTDPKDDKLKPSPAKKPEKAPEPAAPKKWKPVWDDDPDEPEADFTVPAPSKKPDTE 1178

Qy     923 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPT 982

Db    1179 DPADPLG---------GPKTKDPK--------LNKKAPAEKPTEK--------PKPKEVS 1213

Qy     983 KAPKKPTSTKKPKTMPRVRKPK------------TTPTPRKMTSTMP 1017

Db    1214 KEPPKPTEPPKP-AAPKKWKPPWEDDPDEPEADFTMPAPKKPDTEDP 1259
```

RESULT 5
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a mus
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10

```
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor

  Query Match            11.3%;  Score 828.5;  DB 2;  Length 2187;
  Best Local Similarity  28.4%;  Pred. No. 1.5e-27;
  Matches  313;  Conservative  126;  Mismatches  436;  Indels  229;  Gaps   52;

Qy     109 NPTSPPSSKKAPPPS-GASQTIKS----------TTKRSPKP--PNKKKTKKVIESEEIT 155
           :|| ||||| || || ||  : |           ::|: |    |: :|| :|  |  |: 
Db     746 SPT-PPSSKGAPVPSTGAPPSPKGAPIVPTESSISSKQVPAEILPSPQKTPEVTASRLIS 804

Qy     156 EVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK-----VSTSPKIT 210
            |:  | :      ||  |    |              ||| |   |      | ||
Db     805 AVQSPKVDPIMSDVTPTSPKKTSA---------TAVPKDTSATLSLKSVPAVTSLSPPKA 855

Qy     211 TAKPINPRPSLPPNSDTSKETSL--TVNKET------TVETKETTTTNKQTSTDGKEKTT 262
             | |    :|    || : :|     |||       | :     | |  |  |    |
Db     856 PVAPSNEATIVPTEIPTSLKNALAAATPKETLATSIPKVTSPSPQKTPKSVSLKGAPAMT 915

Qy     263 SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKE 322
           | | |   |  ::||::|:     |  |  :     |  : || |   |     :: | : 
Db     916 SKKAT---EIAASKDVSPSQ--FPKEVPLLQHV---PPTSPPKSPVSDTLSGALTSPPPK 967

Qy     323 PTPTTIKSAPTTPKE-PAPTTTKSAPTTPK---------EPAPTTTKEPAPTTPKEPAPT 372
             | |:   || ||: | |  :|  | ||          |  | : | |    ||| : |
Db     968 GPPATLAETPTYPKKSPKPAASKKTPATPSPEGVTAVPLEIPPCSKKAPKTAAPKESSAT 1027

Qy     373 TTKEPAPTT--TKSAPT--------------------TPKEPAPTTPKK-PAPTTPKEP 408
           :: : || |  :|  |:                    || | : ::|: |    |||
Db    1028 SSSKRAPKTAVSKEIPSKGVTAVPLEISLPLKETSKSATPGEKSASSPKRSPKTAGPKE- 1086

Qy     409 APTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTT--PKEPAP----TTPKE 462
             | |   |   |:   |  : |   || |    : :| :| |  |||  |      | |
Db    1087 --TPPGGVTAVPPEISLPPKETPQNATPNESLAASSQKRSPKTSVPKETPPGGVTAMPLE 1144

Qy     463 PAPTTTKEPSPTTPK-----EPAPTTTKSAPTTTKEPAPTTT-KSAPTTPK--------E 508
                 | |    ||     | | |    :| : |: : |   | || ||          |
Db    1145 IPSAPQKAPKTAVPKQIPTPEDAVTILAGSPLSPKKASKTAAPKEAPATPSVGVIAVSGE 1204

Qy     509 PSPTTTKEPAPTTPKEPAPTTPKKPAPTT--PKE------------PAPTTPKEPAP--- 551
             ||:  |    ||| : | | | :| |  |||            |:  :|   | |
Db    1205 ISPSPKKTSKTAAPKENSATLPPKRSPKTAAPKETPATSSEGVTAVPSEISPSPPTPASK 1264

Qy     552 ----TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK-PAPTTPEELA 606
               | | || |  | :| :|||:   |   : | |||      :|:| |    |:| :
Db    1265 GVPVTLTPKGAPNALAE-SPASPKKVPKTAAPEETSTTP------SPQKIPKVAGPKEAS 1317

Qy     607 PTTPEEPTPTT--------PEEPAPTTPKAAAPNTPKEPAPTTPKE-PAPTTPKEPAPTT 657
             | | : || |         | |     |    |:  | |    ||| |||:   | | |
Db    1318 ATPPSKKTPKTAVPKETSAPSEGVTAVPLEIPPSPRKAPKTAAPKETPAPS--PEGATTA 1375

Qy     658 PKETAPTTPKGTAPTTLKEPAPTTP--------------------KKPAPKELAPTTTKE 697
           | :  |:  ||:     ||  ||||                    |  :|||   | : :
Db    1376 PVQIPPSPRKGSKKAGSKE-TPTTPSPEGVTAAPLEIPISSKKTSKMASPKETLVTPSSK 1434

Qy     698 PTS------TTSDKPAPTTPKGTAPTTPKEPAPTTPKE-PAPTTPKGTAPTTLKEPAPTT 750
             |       || : |    |:  | |    ||: |   :||  ||||| | :|::
Db    1435 KLSQTVGPKETSLEGATAVPLEIPPSHKKAPKTVDPKQVPLTPSPK-DAPTTLAE-SPSS 1492

Qy     751 PKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETP--- 807
           ||| |||  || : :  |:   :||| |  | :  |  | | |   :: | :: |||
Db    1493 PKK-APKTAAPPSER-VTTVPPEKPA-TPQKASGTTASKVPVPAETQEVAVSSRETPVTP 1549

Qy     808 --PPTTSEVSTPTTTK------EPTTIHKSPDESTPELSAEPTPKALENSPKE-PGVPTT 858
             ||  :  |   |:|       | |:  || :|    |::  |:   ::||| |  |:
Db    1550 AVPPVKNPSSHKKTSKTIELKEAPATLPPSPTKSPKIPSSKKAPRT--SAPKEFPASPSI 1607

Qy     859 KTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVT 918
           | |  |   :  ||     :|    | |:   |||  :       : |:  :   |: :
Db    1608 K-PVTT--SLAQTAPPSLQKAPSTTIPKENLAAPAV----LPVSSKSPAAPARASASLSP 1660

Qy     919 STTTQDTTPPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKP 978
           :|    | | : ||: :   |    |  :| |   :   |:||::     : ||     |
Db    1661 ATAAPQTAPKEATTIPSCKKAAATETPIETSTAPSLEGAPKETSE----TSVSKVLMSSP 1716

Qy     979 QKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 1038
                ||| :|:|:  |:|                :|:| |   | ::         |
Db    1717 ------PKKASSSKRASTLP-------------ATTLPSLKEASVLS---------PTA 1747

Qy    1039 TPNSKLVEVNPKSEDAGGAEGETP 1062
           | : |   ::| | ||       ||
Db    1748 TSSGKDSHISPVS-DACSTGTTTP 1770


RESULT   6
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

  Query Match            10.9%;  Score 800;  DB 2;  Length 1188;
  Best Local Similarity  28.7%;  Pred. No. 1.3e-26;
  Matches  246;  Conservative   64;  Mismatches  401;  Indels  146;  Gaps   36;

Qy      82 RECDCDAQCKKYDKCCPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPN 141
           |  ||              :  ||    |   | |   |    ||  : |       ||:
Db     397 RPVDCS------------KHVCAGYPTPGGGPPSSPVPGKPAASAPMPS----PHTPPD 439

Qy     142 KKKTKKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHN 201
                 :  | : |           |||  |  |           |:   ||
Db     440 -------VSPEPLPEPSPVPAPAPMPMPTPHSPPAD---------DYVPPTPPVPGKSPP 483

Qy     202 KVSTSPKITTAKPINPRPSL----PPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDG 257
           | ||::       | |||     || :                              |
Db     484 ATSPSPQVQPPAASTPPPSLVKLSPPQAPVG------------------------SPPP 518

Qy     258 KEKTTSAKETQSIEKTSAKDLAPTSKV-----LAKPTPKAETTTKGPALTTPKEPTPTTP 312
           ||||          |     | | |     :  | | |   :  |   :|  | |
Db     519 PVKTTSPPAPIG----SPSPPPPVSVVSPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVAS 574

Qy     313 KEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPA 370
           |   :|  | || : | |   | | |    ::|  |    | ||    | |  |   :
Db     575 PPPPVKSP--PPPTLVASPPPPVKSPPPPAPVASPPPPVKSPPPPTPVASPPPPAPVASS 632

Qy     371 PTTTKEPAPTTTKSAPTTP-KEPAPTTPKKPAPTTPKEPAPTT------PKE---PTPTT 420
           |   | | | |  |:|  | | | |  | |  |   : | | |      | |   | ||
Db     633 PPPMKSPPPPTPVSSPPPPEKSPPPPPPAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTL 692

Qy     421 PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA 480
           | |  |   |:|| :| |::|:||:|  ||||  : |: |       | |     |
Db     693 IPSPPPQEKPTPPSTPSKP-PSSPEKPSP--PKEPVSSPPQTP----KSSPPPAPVSSPP 745

Qy     481 PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKE 540
           ||   | |      :| : ||:|     |:| ::  |||     | |     | |  ||
Db     746 PTPVSSPPALAPVSSPPSVKSSP----PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKS 800

Qy     541 PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAP----TTP----KKLTPTTPEKLAPT 592
           |  |     |   |  | ||   | || : |    ::|    |   | |  |   |
Db     801 SPPLAPVSSPPQVEKTSPPPAPLSSPPLAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPL 860

Qy     593 TPEKPAP-----TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPA 646
           ||:  :|      : || : |:||  | |||   | |: ||:: | ||   | |      |
Db     861 TPKPASPPAHVSSPPEVVKPSTP--PAPTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPP 917
```

```
Qy     647 PTTPKEPAPTTPKETAPTTPKGTAPTTLKE---PA-----PTTPKK---PAPKELAPTTT 695
Db     918 PAMVSSP-PMTPKSSPPPVVVSSPPPTVKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEV 976

Qy     696 KE---PTSTTSDKPAPTTPKGTAPTTP-KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 751
Db     977 KSSPPPTPVSSPPPA---PKSSPPPAPMSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAP 1033

Qy     752 KKPAPKELAPTTTKGPTSTTSDKPAPT-TPKETAPTTPKEPAPTTPKKPAPTTPETPPPT 810
Db    1034 VSSPP---PPVKSPPPPAPVSSPPPPVKSPPPPAPISSPPPPVKSPPPPAPVS--SPPPP 1088

Qy     811 TSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTT 870
Db    1089 VKSPPPPAPVSSPPPPIKSPPPPAPVSSPPPAP------VKPPSLP-PPAPVSSPPPVVT 1141

Qy     871 TAKDKTTERDLRTTPET 887
Db    1142 PAPPKKEEQSLPPPAES 1158
```

```
RESULT   7
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR019c
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009; MIP
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacchar
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseudohy
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosidase
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>
```

```
  Query Match             10.9%;  Score 795.5;  DB 1;  Length 1367;
  Best Local Similarity   27.4%;  Pred. No. 2.3e-26;
  Matches  318;  Conservative  111;  Mismatches  525;  Indels  207;  Gaps   51;

Qy      85 DCDAQCKKYDK---CCPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPN 141
Db     174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSES--STTTSS----- 226

Qy     142 KKKTKKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHN 201
Db     227 --------TSESSTTTSSTSESST------------------------TTSSTSESSTS 253

Qy     202 KVSTSPKI-TTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEK 260
Db     254 SSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPCTKKKTTTSK-TCT---KK 309

Qy     261 TTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPALTTPKE------PTPTTP 312
Db     310 TTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPVTSSTTESSSAPVPTPSSS 362

Qy     313 KEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKE 368
Db     363 TTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVTSS----TTESS 418

Qy     369 PAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTP---TTPKEPA 425
Db     419 SAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSA 471

Qy     426 PTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEPAPTTTKEPSPTTPKEPAP 481
Db     472 PVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSSAPVPTPS-SSTTESSSAP 526

Qy     482 TTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP----TTPKEPAPT-TPKK 532
Db     527 APTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTPVTSSTTESSSAPVPTPSS 586

Qy     533 PAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPT 592
Db     587 STTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSAPVT----SSTTESSSAP- 637

Qy     593 TPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKAAAPNTPKEPAPTTPKEPA 646
Db     638 VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVT----SSTTESSSA 693

Qy     647 PTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDK 705
Db     694 PVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPS-SSTTESS 748

Qy     706 PAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAP 761
Db     749 SAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSVAP 808

Qy     762 TTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKKPAPTTPETPPPTTSEVST 816
Db     809 VPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESSSAPVSSSTTESSVAPVPT 868

Qy     817 PT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTT 871
Db     869 PSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYPGSQTETSVSSTTETTIVP 924

Qy     872 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT--TTQVTSTTTQDTTPFK 929
Db     925 TKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTTVPTTTTTSVTTSSTTTIT 984

Qy     930 ITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDRATNSKATTPKPQKP---- 981
Db     985 TTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--------SESTTTSPTTPVTTV 1036

Qy     982 -------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEA 1028
Db    1037 VSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP--SVTTVTNFTPTT-ITTT 1093
```

```
Qy   1029 MLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQ 1088
Db   1094 VCSTGT-------------NSAGETTSGCSPKTVTTTVPCSTGTGEYTTEATTLVTTAVT 1140

Qy   1089 GIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGI 1148
Db   1141 TTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTLAPSAPVTPATN-----AV 1188

Qy   1149 PSPIDTVFTRC----NCEGKT 1165
Db   1189 PTTITT--TECSAATNAAGET 1207


RESULT   8
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of chick
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1


  Query Match             9.6%;  Score 704.5;  DB 2;  Length 1151;
  Best Local Similarity   25.7%;  Pred. No. 1.2e-22;
  Matches  304;  Conservative  104;  Mismatches  445;  Indels  331;  Gaps   59;

Qy    105 AEVHNPTSPPSSKK------APPPSG-----------------ASQTIKSTTKRSPKPP 140
Db     39 AELH-PAAPQPPPKWVPIGGAPPPPGTEPTPPSKPTDGADAAPKASAELTSPPPASPSPP 97

Qy    141 NKKKTKKVIESEEITEVKDNKKNRTKKKPTPKP-------PVVDEAGSGLDNGDFKVTTP 193
Db     98 DGPKAPSGAGEAE-------------AGTPPPSQGPAGTPPPSQGAAGAPKGDGTAQPS 143

Qy    194 DTSTTQHNKVSTS--PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNK 251
Db    144 GTKSGADGKPAAQDVPKATTA-ATEARP--------ASAASPTVPKATAEATAVTAASQS 194

Qy    252 --QTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTP 309
Db    195 APKAATDAAAVTAA---SQSAPKATV-EVKPAAAAVAKEAKAVTAAAAAPKATAEAKPAP 250

Qy    310 -TTPKEPASTTPKEP----TPTTIKSAPTTPKEPAPTT----------TKSAPTTPKEP 353
Db    251 VTSPTIPCSSAEAKPLTAASPTASKA--TAEAKPVPATASLMATKVTAEAKPAPSPSVPK 308

Qy    354 APTTTK----------------------EPAPTTPKEPAPTTTKEPAPTTTKSAPTT-- 388
Db    309 ATTDTKAVTATAPKAGPDVKPAVAVCAEAKPAPPPPPQQLPKAAAAAAPTGTELKPATAP 368

Qy    389 ----PKEPA-------PTTPKKPAPTTPKEPAPTTPKEPTPTT----PKEPAPTTKEPAP 433
Db    369 PHGSPRANSHTVTVTPPNVPRAAAATVP--TAGAVPKASTGTTPAAAPQQPVP---KAAP 423

Qy    434 TTPKEPAPTAPKKPAPTTPKEPAPTTPKEP---APTTTKEPSPTTPKEPAPTTTKSAPTT 490
Db    424 VTPPSPQQAVPRAATAAA----APVTPQQPVTKAATTTNATPPPQPIPKAATTTTATPVT 479

Qy    491 TKEPAP---TTTKSAPTTPKEPSPTTTKEP----APTTPKEPAPTTPKKPAPTTPKEPAP 543
Db    480 PQQPIPKAGTDAAPPPAVPKAPSDGRAATPGVPNAATDPQKPPPTPQSVPSAVTEPKPQP 539

Qy    544 ---------TTPKEPAPT-TTKKPAPTAPK-------EPAPTTPK---ETAPTTPKKLTP 583
```

```
Db    540 RAAPPPSNEATPAVPSPSPNLKSPLPTIPKPVPLMALTPQPVTAQMVTQLAATKPSPIVP 599

Qy    584 TTPEKLAPTTPEKPAPTTPEELA-------PTTP----------EEPTPTTP----EEPA 622
Db    600 KASPK-ALMTPPPPPPGLPRALAAAKLLGLPSSPVASAMHAKVTPRPLPASPVPMAASPA 658

Qy    623 PTTP---------KAAAPNTPKE----------PAPTTPKEPAPTTPKEPAPTTPKETAP 663
Db    659 SLGPDAARVALATNAASPGAKPEAAGGNGTLMAPMGAANTQMAPIGAAGAAQTAPMGAAH 718

Qy    664 T-------------TPKGTAPT-------------------TLKEPAPTTPKKPAPKEL 690
Db    719 THVSPMGAGGATQMSPTGAANTHMSPIGAGGATQMSPMGAANTQMSPMGATTTQMSPMGA 778

Qy    691 APTTTKEP--------TSTTSDKPAPTTPKG--TAPTTPKEPAPTTPKEPAPTTPKGTAP 740
Db    779 AATTQPSPMGAAATQVTATSAGNTMQVSPMGAATPPQTPSVGAATTP-QPSPM----GAA 833

Qy    741 TTLKEP--APTTPKKPAPKELAPTTTKGP----TSTTSDKP-APTTPKET---APTTPKE 790
Db    834 TTLMSPMGAATTPQ---PSPMGAVTTQPPPMAATNTTQPPPMAASTPQSTPMGAATTTQS 890

Qy    791 P---APTTPKKP-APTTPETPP-------PTTSEVSTPTTTKEPTTIHKSPDESTPEL-- 837
Db    891 PPMGATTTQSPPMGASTPQAPPTVAGSPTPPPPIPPSPTAQTSPQPMSKSPPPDPPKAPS 950

Qy    838 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMT-K 896
Db    951 AAAQTSPAAHVANASPGV-TAVSPA---PIGVTEASPSADGARLSPGPTAATDGPKASPA 1006

Qy    897 ETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMN 956
Db   1007 ATADVTEAATD--VTAAATAVPA----EAAP---------TKAKRSSSSSSSSSSSSSS 1051

Qy    957 KPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTM 1016
Db   1052 SSSSSSSSDSDSSSSSSESNPASPA--------------PAVGDGQQQMTPGAAQSVP 1096

Qy   1017 PELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGE 1060
Db   1097 P-------VTEAAVQ-------------EAAAAAAAAAGAERE 1119


RESULT   9
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3


  Query Match             9.2%;  Score 674;  DB 2;  Length 1229;
  Best Local Similarity   27.8%;  Pred. No. 2.4e-21;
  Matches  328;  Conservative   89;  Mismatches  384;  Indels  380;  Gaps   56;

Qy     73 RCFESFERGRECDCDAQCKKYDKCCPDYESFCAEVHNPTSP---------PSSKKAPPPS 123
Db     34 RDFKIVKHIKKNTCTCSC----KCVPDAPSNPFDVSTTISSINNDNVDIGPSGDSNPTGS 89
```

```
Qy    124 GASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGSGL 183
Db     90 SWFQEIEATVGGQ----TVKSEHNIDSSVEV-----------EKKVT------------- 121

Qy    184 DNGDFKVTTPDTSTTQ---HNKVSTSPKITTA-KPINPRPSLPPNSDT-------SKETS 232
Db    122 -------TSTDASTTNAPTTGKDSTTPEIITGIVVINSKSESVTDMSTTRFSTTLSPTTE 174

Qy    233 LTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKA 292
Db    175 LLTSPETLVSTDSSTST-EQTSPDNTTEIASPMETNTTTEATTTSVEPSVSTLAS---ED 230

Qy    293 ETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKE 352
Db    231 ETTVTAIAEST------TTVIAEVSTTTEEPTTT-----------AESTTKKSTT---- 268

Qy    353 PAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----KKPAPTTPKEP 408
Db    269 KAPATTEEPTPTTTEE--VTTTEAETSTTTSSETSTEK---PTTPLIDNKIAGPATGK-- 321

Qy    409 APTTPKEP-TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTT 467
Db    322 -PETTHFPVTGTTPN--FDTATETPFVAKSEDKMTLSKTAATETTQQTTEVT-DGPEKET 377

Qy    468 TKE-----PSPTTPKEPAPTTTKSAP-------TTTKEPAPT----TTKSAPTTPKEPSP 511
Db    378 TKNVSIEIPITTVPLVETTSTTSTASKESDGFHTTLKLKVTTADSDSTESATTVKPFNEE 437

Qy    512 TTTKE---PAPT-------TPK------EPAPTTPKKPAP---------------------- 535
Db    438 TTTKSHVVPKPTKKGTVKVTPKLELSFDEPTEIT-KAPHPGKLLEKKTYHFVLSDNFARY 496

Qy    536 TTPKE---------------PAPTTPKEPAP----TTTKKPAPTA--PKEPAPTT---PK 571
Db    497 SEAKENDDYNHLDYNHYREAKEPTTTEESSTTEEVTTTEEPANTGNPPTTENPTTTEQPT 556

Qy    572 ETAPTTPKKLTPTTPEKLA---PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKA 628
Db    557 STAESTTTALPFTTEQTVTTEEPTTAEKSTATQ----KPTTTQESVST---EKTSTTKKA 609

Qy    629 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK 688
Db    610 ---STTEE--PTTTDEPTTTT---ESSTTGKATTPELSTTSEETTTTELKITTE------ 655

Qy    689 ELAPTTTKEPTSTTSDKPAP----TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLK 744
Db    656 --GSTTTEEPTTTAIFAEASTGIITTDEETTSTTSTTPEITSTKE--IVTESAITQTSVS 711

Qy    745 EPAPTTP----------------------KKPAPKELAPTTTKGP-------------- 767
Db    712 VVESSTPRQLPERWKAIVNKFKHNLEVLKEKKRLLKEKESTSTTGSDSSETTTVVAENID 771

Qy    768 --TSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPE--TPPPTTSE-VSTPTTTKE 822
Db    772 EVTTTEKEKVVQTTPITTEKSTTQEETTTTTTTTEKTTSKTTTEKPTTSESATTETTTSE 831

Qy    823 PTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLR 882
Db    832 PST-----TEST----------------------TVDTSSATTEESSTAAETTTTSAE-- 862

Qy    883 TTPETTTA-------------------------------------APKMTKETATTTE 903
Db    863 -TSETTTSESAAFITGESPENTALQSSSQKSEENESSAEKPGARRDFVPKKHKTTVKPAE 921

Qy    904 KTTESKITATTTQ-VTST---TTQDTTPFKITTLKTTTLAPKVT---TTKKTITTTEIMN 956
Db    922 TTSAVAASTTTTEPITTTEKSTTLETTPIEATTLNEVTGPAFVTGAPVDETTINTLELLS 981

Qy    957 KPEETAKPKDRATNSKATTPKPQ----------------KPTKAPKKPTSTKKPKTMPR 999
Db    982 K----------INNTQISQPKPTDISKTDALSSLISGLIGSFTKAPMAPTI--------- 1022

Qy   1000 VRKPKTTPTPRKMTSTMPELN-----------PTSRIAEA 1028
```

```
Db   1023 ----HTTTDAAFVTATEASLNDGSDKKIIDEAQPTDEIRRA 1059
```

RESULT 10
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episia
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precurso
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218;
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129;
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burc
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epith
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated se
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky,
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may gen
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associa
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu

```
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut agglu
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokine r
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region ar
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 125 c
 partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively gly
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48146
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice form
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predicted
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

  Query Match             9.2%;  Score 672;  DB 1;  Length 1344;
  Best Local Similarity   27.8%;  Pred. No. 3.2e-21;
  Matches  310;  Conservative   78;  Mismatches  472;  Indels  256;  Gaps   56;

Qy    233 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 289
Db     15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy    290 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTP----------KE 322
Db     75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy    323 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 362
Db    135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy    363 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 408
Db    191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy    409 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 445
Db    248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy    446 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 498
Db    308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy    499 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 547
Db    368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy    548 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 603
Db    425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy    604 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 643
Db    485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy    644 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 690
Db    545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy    691 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 744
Db    601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy    745 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 788
Db    656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy    789 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 831
Db    715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy    832 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 880
Db    773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy    881 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 931
Db    833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy    932 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 989
Db    889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy    990 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 1041
Db    947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy   1042 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 1101
Db   1004 AHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy   1102 ICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1137
Db   1050 ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 1085


RESULT  11
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31

  Query Match             9.1%;  Score 670;  DB 2;  Length 7962;
  Best Local Similarity   24.4%;  Pred. No. 2.1e-20;
  Matches  267;  Conservative  111;  Mismatches  471;  Indels  246;  Gaps   50;

Qy     98 PDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRS----------PKPPNKK 143
Db   6827 PEEEEFVPE--EEVLPEVKPKVPVPAPVPEIKKKVTEKKVVIPKKEEAPPAKVPEVPKKV 6884
```

```
Qy    144 KTKKVI---ESE----EITEVKDNKKNRTKKKPTPKPPV--VDEAGSGLDNGDFKVTTPD 194
          : |::|   | |    |:|| : :    :: |   | :  |:|           |   |:
Db   6885 EEKRIILPKEEEVLPVEVTEEPEEEPISEEEIPEEPPSIEEVEEVAPPRVPEVIKKAVPE 6944

Qy    195 TSTTQHNKVSTSP-----KITTAK---PINPRPSLP------PNSDTSKETSLTVNKETT 240
            |    ||   |     ||   |   |:  : : |      |     |:  |   ||
Db   6945 APTPVPKKVEAPPAKVSKKIPEEKVPVPVQKKEAPPAKVPEVPKKVPEKKV-LVPKKEAV 7003

Qy    241 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAK---------------------- 276
             |  |   :: |   :::    ||   :|   |:
Db   7004 PPAKGRTVLEEKVSVAFRQEVV-VKERLELEVVEAEVEEIPEEEEFHEVEEYFEEGEFHE 7062

Qy    277 ------------------------DLAPTSKVLAKPTPKAETTTKGPALT-----TPKE 306
                                   ::    :|     |||    ||| ::       |
Db   7063 VEEFIKLEQHRVEEEHRVEKVHRVIEVFEAEEVEVFEKPKA--PPKGPEISEKIIPPKKP 7120

Qy    307 PTPTTP-KEPASTTPKEPTPTTIKSAPTTPKEP-APTTTKSAPTTPKEPAPTTTKEPAPT 364
          ||   | ||| :  |: |    ::     |:|| | |       |||  |
Db   7121 PTKVVPRKEPPAKVPEVPKKIVVEEKVRVPEEPRVPPTKVPEVLPPKEVVP--------- 7171

Qy    365 TPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKK----PAPTTPKEPAPTTPKEPTPTT 420
            | | |   |  ||      |  |: |    |:|    | |  |: |    |: |
Db   7172 EKKVPVPPAKKPEAP-----PPKVPEAPKEVVPEKKVPVPPPKKPEVPPTKVPEVPKAAV 7226

Qy    421 PKEPAPTTKEPAPTTP-------KEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSP 473
          |::  |    | | :|       | :|:|| |  |  |    |  |||  |      :|
Db   7227 PEKKVPEAIPPKPESPPPEVFEEPEESPSAPPK-KPEVPPVRVPEVPKEVVPEKKVPAAP 7285

Qy    474 TTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKP 533
            ||:|  |  |  |   ||  |       |  ||:|    ||       |: |    |:|
Db   7286 --PKKPEVTPVK-VPEAPKEVVPEKKVPVP-PPKKPEVPPTK-----VPEVPKVAVPEKK 7336

Qy    534 APTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLT----PTTPEKL 589
            |    |  |  |: | |   ::|   | :||     :|  |  | ::|       |||
Db   7337 VP----EAIPPKPESPPPEVFEEPEEVALEEPPAEVVEEPEPAAPPQVTVPPKNPVPEKK 7392

Qy    590 APTT----PEKPAPTTPEELAPTTPEEPTP-TTPEEPAPTTPKAAAPNTPKEPAP----T 640
            ||       || |    ||       ||:  |   |::  |  | |  |  |||  |
Db   7393 APAVVAKKPELPPVKVPEVPKEVVPEKKVPLVVPKK--PEAPPAKVPEVPKEVVPEKKVA 7450

Qy    641 TPKEP------APTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTT 694
            ||:|       |  ||:|      |  |  |    |   : | |   ::  |:|:||
Db   7451 VPKKPEVPPAKVPEVPKKPV----LEEKPAVP---VPERAESPPPEVYEE--PEEIAP-- 7499

Qy    695 TKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKP 754
            :|  :   :|| |   :      |: | |  |:||    |:   |  :|:|    ||:|
Db   7500 -EEEIAPEEEKPVPVAEE----EEPEVPPPAVPEEPKKIIPEKKVP-VIKKPEAPPPKEP 7553

Qy    755 APKELAPTTTKGPTSTTSDKPAPTTPKETA-------PTTPKEPAPTTPKKP--APTTPE 805
            |:::     |         | |  |||           ||:  |  |: |     ||
Db   7554 EPEKV---IEKPKLKPRPPPPPPAPPKEDVKEKIFQLKAIPKKKVPENPQVPEKVELTPL 7610

Qy    806 TPPPTTSEVS--TPTTTKEP-------TTIHKSPDESTPELSAEPTPKALENSPKEPGVP 856
            |    :|    |    ||       : : | |:|  |::     || ||   |:| ||
Db   7611 KVPGGEKKVRKLLPERKPEPKEEVVLKSVLRKRPEEEEPKVE----PKKLE-KVKKPAVP 7665

Qy    857 TTKTPAATKP----EMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITA 912
              |   ||    |: |  |  :    ||   | |     |     |
Db   7666 ---EPPPPKPVEEVEVPTVTKRERKIPEPTKVPEIKPAIPLPAPEPKPKPEA-------- 7714

Qy    913 TTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNK------PEETAKPKD 966
                      :: |:|   : |: |     :|   :  |       || ||||
Db   7715 ----------------EVKTIKPPPVEPEPTPIAAPVTVPVVGKKAEAKAPKEEAAKPKG 7758

Qy    967 RATNSKATTPKPQKPTKAPKKPTS--TKKPKTMPRVRKPKTTPTP-----RKMTSTMPEL 1019
                  || | :  :   :| |   | |   |   : |  |         : :  |  |
Db   7759 PIKGVPKKTPSPIEAERRKLRPGSGGEKPPDEAPFTYQLKAVPLKFVKEIKDIILTESEF 7818

Qy   1020 NPTSRIAEAMLQTTT 1034
            :| | |  ::  :|
Db   7819 VGSSAIFECLVSPST 7833
```

```
RESULT 12
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/
 3504/1


  Query Match             9.1%;  Score 668;  DB 2;  Length 3507;
  Best Local Similarity   22.9%;  Pred. No. 1.2e-20;
  Matches  282;  Conservative  175;  Mismatches  480;  Indels  296;  Gaps   51;

Qy     27 LSSCAGRCGEGYSRD-ATC-------NCDYNCQHYMECCPDFKRVCTAE----------- 67
          : ||  :|  ||: | |||         |        |  :: |:|  |
Db   1781 MGSCGCKCMAGYTGDGATCIKIEEEEPKSDKTA-----CTDEWSRLCELEKKQCTVDEEEV 1835

Qy     68 -------------------LSCKGRCFESFERGRECDC-----DAQ-------------- 89
                             |   | | :  :  :  :|     |:
Db   1836 PQCGACLPGHHPINGTCQSLQISGLCAQKNDCNKHAECIDIHPDSHFCSCPDGFIGDGMI 1895

Qy     90 CKKYDKC-----CPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKK 144
          |   |:|     | |  : |       :   |       |  :  :       | ||:|
Db   1896 CDDVDECNNAGMCDDENTKC------ENTIGSFNCVCLEGFKKVDEKCVVDEKKQPNREK 1949

Qy    145 TKKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVS 204
                || |   :  |  ::||| |  |   :             |   |||    |:
Db   1950 I-------EIDEENSSSSNSGQEKPTTKGIVSSTS----------ATSSESTTAEPHVT 1991

Qy    205 TSPKITTAKPINPRPSLPPN---SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEK 260
          ||   ||:         | |   |  | | | : :| ||  ||  ::| :: :|::
Db   1992 TSISSTTSTKDMTSSKSPENVTMSSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLT 2051

Qy    261 TTSAKETQSIEKTSAKDLAP----TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPA 316
          :: |  |: | ::| |    |    :|::  | : |:   |:    ::|  |: || :
Db   2052 SSPATTTEVITESSVKSTTPKEESSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVT 2110

Qy    317 STTPKEPTPTTIKS-APTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTK 375
          || |:    | : | || |   |    | |: | |   | :|| :   |||   :  : |
Db   2111 STVPETSKSTVLSSEAPVTSTSPTEVHT-SSETKPSLSASSTTGDTNSTTPSTSSLASVK 2169

Qy    376 EPAPTTTKSAPTTPKEPAPTTPKKPAPTTPK-EPAPTTPKEPTPTTPKEPAPTTKEP-AP 433
             :    ||   | : :  :|    |:|   :   ::  : : |:       :| || :
Db   2170 STSAPEGTSASVAPVKLSSLSPDVSQPSTKTFDATESSTVQASETSSGTSVKSTSEPESH 2229

Qy    434 TTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP--KEPAPTTTKSAPTTT 491
          |       : |    | |  :  || |:     :|::|: |||  :   |  : |   ||
Db   2230 VTKLSITSSNPSSSVPVTSPKSTPTVPE-----STEQPTSTTPSGQSLTPMNSNSEVLTT 2284

Qy    492 KEPAPTTTKSAP------TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTT 545
          ||   ::  :|       |||   | ::| |    |||  :  :      |:| : |   :
Db   2285 SEPHVLSSSLSPDVSQSSTTPNNLSESSTVE----TPKTSSEVSLNSEEPSTTEAPTTLS 2340

Qy    546 PKEPAPTT--------------------TKKPAPTAPK----------EPAPTTPKETA 574
          |   : ||                    ::||  :||:           :|  | |:
Db   2341 PDILSTTTNNLSQSSTVSTEDRSEISSENSEKPT-SAPELVTSSVTHVASSSPDVPTES- 2398

Qy    575 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP----------AP 623
```

```
Db   2399 -SEPDDLTGSSTENIPEASSKQTISSTPTPDTTTASEEPTKSTSMSPDLSTTSNVLSESS 2457

Qy    624 TTPKAAAPNTPKEPAPTT------------PKEPAPT---------TPKEPAP-----TT 657

Db   2458 TTPESSS----KSPVSSSTEGISVVTSTEFSKVPESTISSVLEEDLTKTTPSPILEETTT 2513

Qy    658 PKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPA---------P 708

Db   2514 ASETSEPLTEDSLTVSVRIHELTTSSENVPKESESTTTSSESSKPSQEPAGILTSTVVVP 2573

Qy    709 T-----------------TP--KGTAP--TTPK-------EPAPTTPKEPAPTTPKGTAP 740

Db   2574 TSSVSLITASEIEAITSNTPFKQGRTPITTSPKSLVKSTTSPSTVTSSEPSESTKRTTVS 2633

Qy    741 TTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDKPAPTTPKETAPTTPKEPAPTTPKKP 799

Db   2634 TTVSTTTPTEETTTSESLILTAAPSKPTESTTESSEAPTTPAKTSETKPSNVSSTSRKST 2693

Qy    800 APTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP--ELSAEPTPKALE--NSPKEPGV 855

Db   2694 ENVETSTSQSGSLESSTMSSTS------SEPETNAPAVTVSSEASSTTLEENSSTSSP-- 2745

Qy    856 PTTKTPAATK-----PEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKI 910

Db   2746 --TSSEASVKLSSLFPESITSEAVTVSSR----APAEITMSSESHREISTVSSEPSEPEI 2799

Qy    911 TATTT--------------QVTSTTTQDTTP-FKITT-----LKTTTLAPKVTTTKKTI 949

Db   2800 PLSTTVSPNVVTASSIPSEEPILSSVTSSSTPRVRLITGTPDDLIVSVTVPSHGNRRQNI 2859

Qy    950 TTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK---TT 1006

Db   2860 TASSV---PSNSTSPIILPSES-LTTPQPPPTTTTTAKPATTSGKRGPPSIQPPAEMFTT 2915

Qy   1007 PTPRKMTSTMPELN-----PTSRIAEAMLQTTT 1034

Db   2916 PAP------PPPSNGGYGEETNQEEEQVTSTTT 2942
```

RESULT 13
T45025
mucin MUC5B, tracheobronchial [imported] - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 21-Jan-2000 #sequence_revision 21-Jan-2000 #text_change 21-Jul-2000
C;Accession: T45025
R;Desseyn, J.L.; Guyonnet-Duperat, V.; Porchet, N.; Aubert, J.P.; Laine, A.
J. Biol. Chem. 272, 3168-3178, 1997
A;Title: Human mucin gene MUC5B, the 10.7 kb large central exon encodes various alternat
A;Reference number: Z22899; MUID:97166151
A;Accession: T45025
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3570 <DES>
A;Cross-references: EMBL:Z72496; NID:g1834502; PIDN:CAA96577.1; PID:g1834503
A;Experimental source: placenta
C;Genetics:
A;Gene: MUC5B

Query Match 9.1%; Score 663; DB 2; Length 3570;
Best Local Similarity 26.1%; Pred. No. 1.9e-20;
Matches 322; Conservative 92; Mismatches 489; Indels 330; Gaps 54;

```
Qy      54 MECCPDFKRVCTAELSCKGRCFESFERGRECDCDAQCKKYDKCCPDYE--SFCAEV-HNP 110

Db   2279 VECSLDFGLVC--------------RNRE-----QVGKF-KMCFNYEIRVFCCNYGHCP 2317

Qy     111 TSPPSSKKAPPPSGASQT---IKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKK 167

Db   2318 STPATSSTATPSSTPGTTWILTKLTT----------TATTTESTGSTATPSS------- 2359

Qy     168 KPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAK--------PINPRP 219

Db   2360 --TQGPP----AG----------TPHVSTT-----ATTPTVTSSKATPFSSPGTATALP 2397

Qy     220 SLPPNSDTSKETSLTVNKETTVET-----KETTTTNKQTSTDGKEKTTSAKETQSIEKTS 274

Db   2398 ALRSTATTPTATSFTAIPSSSLGTTWTRLSQTTTPMATMSTATPSSTPETVHTSTVLTTT 2457

Qy     275 AKDLAPTSKVLAKP--TPKAETTTKGPALTT-----PKEPTPTTPKEPASTTPKEPTPTT 327

Db   2458 ATTTGATGSV-ATPSSTPGTAHTTKVPTTTTTGFTVTPSSSPGTARTPPVWISTTTTPTT 2516

Qy     328 IKSAPTTPKEPAPTTTKSAPTTPKEPAPT-TTKEPAPTTPKEPAPTTTKEPAPTTTKSAP 386

Db   2517 SGSTVTPSSIPGTTHTPTVLTTTTQPVATGSMATPSSSTQTSGTPPSLITTATTITATGS 2576

Qy     387 TT-PKEPAPTTPKKP-APTTPKEPAPT----TPKEPTPTTPKEPAPTTKEPAPTTPKEPA 440

Db   2577 TTNPSSTPGTTPIPPELTTTATTPAATSSTVTPSSALGTTHTPPVPNT--TATTHGRSLS 2634

Qy     441 PTAP-------------------KKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA 480

Db   2635 PSSPHTVRTAWTSATSGTLGTTHITEPSTGTSHTPAATT----GTTTTSTPALSSP-HPS 2689

Qy     481 PTTTKSAPTT-TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPK 539

Db   2690 SRTTESPPSPGTTTPGHTTATSRTTATATPSKTRTSTLLPSQPTSAPITT----VVTTGC 2745

Qy     540 EPAPTTPK-------EPAPT--------------------------TTKKP-------- 557

Db   2746 EPQCAWSEWLDYSYPMPGPSGGDFDTYSNIRAAGGAVCEQPLGLECRATAQPGVPLGELG 2805

Qy     558 ------------------------------------APTAPKEPAPTTPKETAPTT- 577

Db   2806 QVVECSLDFGLVCRNREQVGKFKMCFNYEIRVFCCNYGHCPSTPATSSTAMPSSTPGTTW 2865

Qy     578 -PKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTP-EEPAPTTPKAAAPNTPK 635

Db   2866 ILTELTTT-----ATTTASTGSTATPSSTPGTAPPPKVLTSPATTPTATSSKATSSSSPR 2920

Qy     636 EPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTT 695

Db   2921 -TATTLPVLTSTATKSTATSVTP------IPSSTLGTTGTLPEQTTTPVATMSTIHPSST 2973

Qy     696 KEPTSTTSDKPAPTTPKGTAPTTPKEPAP-----TTPKEPAPTTPKGTAPTTLKEPAPTT 750

Db   2974 PETTHTSTVLTTKATTRATSSTSTPSSTPGTTWILTELTTAATTTAGTGPTATPSSTPGT 3033

Qy     751 PKKPAPKELAPTTT--KGPTSTTSDKPAPT---TPKETAPTTPKEPAPTTPKKPAPTTPE 805

Db   3034 TWILTELTTTATTTASTGSTATLSSTPGTTWILTEPSTTATVTAPPGSTATASSTQATAG 3093

Qy     806 TPPPTTSEVSTPTTTKEPTTIHKSPDEST--PEL-SAEPTPKALENSPKEPGVPT----- 857

Db   3094 TPHVSTT-ATTPTVTSSKATPSSSPGTATALPALRSTATTPTATSFT----AIPSSSLGT 3148

Qy     858 -------TKTPAATKPEMTTTAKDKT--TERDLRTTPETT-----TAAPKMTKETATTTE 903

Db   3149 TWTRLSQTTTPTATMSTATPSSTPETVHTSTVLTTTATTTGATGSVATPSSTPGTAHTTK 3208

Qy     904 --KTTESKITAT--------------TTQVTSTTTQDTT--------------PFKIT 931

Db   3209 VPTTTTTGFTATPSSSPGTALTPPVWISTTTTPTTTTPTTSGSTVTPSSIPGTTHTARVL 3268

Qy     932 TLKTTTLA----------------PKVTTTKKTITTTEIMNKPEET--AKPKDRATNSK 972

Db   3269 TTTTTTVATGSMATPSSSTQTSGTPPSLTTTATTITATGSTTNPSSTPGTTPIPPVLTSM 3328

Qy     973 ATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQT 1032

Db   3329 ATT-----PAATSSKATSSSSPRTATTL--PVLTSTATKSTATSFTPIPSS----TLWTT 3377

Qy    1033 TTRPNQ--TPNSKLVEVNPKSEDAGGAEGETPH 1063
```

```
Db   3378 WTVPAQTTTPMSTMSTIHTSSTP------ETTH 3404


RESULT  14
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp180
A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNopp180
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein


  Query Match             8.8%;  Score 646;  DB 2;  Length 990;
  Best Local Similarity   25.8%;  Pred. No. 2.9e-20;
  Matches  260;  Conservative  134;  Mismatches  389;  Indels  226;  Gaps   46;

Qy    112 SPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKKKPT- 170
Db     57 SPDAKKRKPPANGL--------------PKKKSAKESSSSEDSSSEEDEPPAKKRAQPAG 102

Qy    171 PKPPVVDEAGSGLDNGDFKVTTPD--TSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTS 228
Db    103 GKKPVV------------KAVQPKKAKSSSEDSSDESDSEEETKKPPAKRPAQTPK---- 146

Qy    229 KETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKP 288
Db    147 ------------VAAVKTPTQKKAKS--------SSSESSSSEDEASKKKQPVIKV---- 182

Qy    289 TPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTPKEPAPTTTK 344
Db    183 PPKQAVVKAGLASNNGKTADSSSSEDSDSPPAKKTAATKTPPT---KPATAAKPQAKKTA 239

Qy    345 SAPTTPKEPAPTTTKEPAPTTPKEPAPTT-TKEPAPTT-----TKSAPTTPKEPAPT--- 395
Db    240 GKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVPPPTSVSKKKTLSQPGTKAKPESSDSS 299

Qy    396 -------TPKKPAPTTPKEPAPTTP----KEPTPTTPKEPAPTTKEPAPTTPKEPAPTAP 444
Db    300 DSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSESDSSSEDEKKSSVKLGVKAAP 359

Qy    445 KKPAPTTPKEPAPTTP----KEPAP----------------TTTKEPSPTTPKEPAPTT 483
Db    360 KK-APAAP--DAKSTPVAAAKKSAPAKKASSSSDSDSSSNEETTTKPAAKTTPAKSAATP 416

Qy    484 TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PTTPKKPAPTTPK 539
Db    417 TSKTPTNGK---ATPTSKTPAKPGTPKTSTAKKDSSSSDSSDSSSDEETTTKPAAKTTPA 473

Qy    540 EPAPT-TPKEPAPTTTKKPAPTAPKEPA-PTTPKETA------------PTTPKKLTPTT 585
Db    474 KSAATPTSKTP---TNSKATPTSKKTPAKPGTPKTSAAKKDSSSSDSSDSSSDEKKTPA- 529

Qy    586 PEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP-KAAAPNTPKEPAPT---- 640
Db    530 -KRAAKTTPAKPA------AKTTPAKPA-------AKTTPAKPAAKSTPGKQVPTKKES 574

Qy    641 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPT 699
Db    575 SSSDSSDSSSEDEKKSSAKPAVKTTPGKATS-------KPVVASKPVPAK------KASS 621

Qy    700 STTSDKPAPTTPKGTAPTTPKEPAPTT--PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPK 757
```

```
Db    622 SSDSDSSEEETTKTTKPLTKLSPAVKTLPPKKAESSSDSSSDSDSEKK---TKPAKPPAK 678

Qy    758 ELAPTTTKGP--------TSTTSDKPAPTTPKETAPT--TPKEPAPTTPKK-PAPTTPET 806
Db    679 SATPVNTKAPAQNKASKASCSDSDSSSEEEGKSKQPTGKSPAAKATAPPKKNPVAVNKDK 738

Qy    807 PPPTTSEVSTPTTTKE-PTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTTKTPAATK 865
Db    739 PSSSSSSDSSGDDEKQKPKQAAAAKDVKQGAKAAKPTPKKAASSSSE---DSSSDEDVSK 795

Qy    866 PEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDT 925
Db    796 AKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGGKNTSTTKIANST---- 847

Qy    926 TPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKDRATNS---KATTPKPQ 979
Db    848 --------------PKAAAAECSEESSSSEDEGKANGTSGKRKRESTGNAECEAVTPE-- 891

Qy    980 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPTSRIAE 1027
Db    892 ------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP--RMAD 932


RESULT  15
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume i
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>


  Query Match             8.6%;  Score 633;  DB 2;  Length 489;
  Best Local Similarity   31.2%;  Pred. No. 5.2e-20;
  Matches  149;  Conservative   33;  Mismatches  252;  Indels   44;  Gaps    7;

Qy    335 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 394
Db     39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy    395 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 451
Db     84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy    452 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 511
Db    142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP-------------PPSPSPPPP 187
```

```
Qy    512 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 571
Db    188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy    572 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 631
Db    248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy    632 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 691
Db    308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYY- 366

Qy    692 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 751
Db    367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy    752 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 809
Db    423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPP 475
```

Search completed: April 26, 2002, 16:32:49
Job time: 764 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:32:55 ; Search time 62.75 Seconds
(without alignments)
795.232 Million cell updates/sec

Title: AA7
Perfect score: 7323
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1361

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1194	16.3	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	950	13.0	1664	1	SLP1_CLOTM	Q06852 clostridium
3	795.5	10.9	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	8.9	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	566.5	7.7	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	7.6	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	7.5	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	7.3	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	7.2	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	7.2	865	1	CPN_DROME	Q02910 drosophila
11	518.5	7.1	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	7.0	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	6.9	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	6.8	1970	1	RPB1_MOUSE	P08775 mus musculu
15	498	6.8	831	1	NFH_RAT	P16884 rattus norv
16	493.5	6.7	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	6.7	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488.5	6.7	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	6.5	267	1	EXTN_MAIZE	P14918 zea mays (m
20	473.5	6.5	2142	1	BAT2_HUMAN	P48634 homo sapien
21	471.5	6.4	5376	1	ZAN_MOUSE	O88799 mus musculu
22	468.5	6.4	634	1	HWP1_CANAL	P46593 candida alb
23	467	6.4	817	1	VRP1_YEAST	P37370 saccharomyc
24	454	6.2	797	1	VGLX_HSVEB	P28968 equine herp
25	454	6.2	1161	1	YJ9P_YEAST	P47179 saccharomyc
26	448.5	6.1	670	1	VG50_HSVI1	Q00130 ictalurid h
27	448.5	6.1	1229	1	N121_HUMAN	Q9y2n3 homo sapien
28	442.5	6.0	751	1	FP1_MYTGA	Q27409 mytilus gal
29	440	6.0	1794	1	YAV1_SCHPO	Q10172 schizosacch
30	439.5	6.0	1083	1	T2D3_HUMAN	O00268 homo sapien
31	438	6.0	2774	1	MAPA_RAT	P34926 rattus norv
32	434	5.9	3164	1	TEGU_HSV11	P10220 herpes simp
33	432.5	5.9	439	1	XP2_XENLA	P17437 xenopus lae
34	431.5	5.9	1125	1	MAP4_MOUSE	P27546 mus musculu
35	431	5.9	3256	1	KI67_HUMAN	P46013 homo sapien
36	428	5.8	1185	1	DRPL_HUMAN	P54259 homo sapien
37	426	5.8	1183	1	DRPL_RAT	P54258 rattus norv
38	426	5.8	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
39	425.5	5.8	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
40	425	5.8	3421	1	TEGU_HSVEB	P28955 equine herp
41	424	5.8	307	1	SGS3_DROME	P02840 drosophila
42	424	5.8	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
43	420.5	5.7	1411	1	TCOF_HUMAN	Q13428 homo sapien
44	419.5	5.7	2476	1	ZAN_PIG	Q28983 sus scrofa
45	418.5	5.7	907	1	VGP3_EBV	P03200 epstein-bar

ALIGNMENTS

RESULT 1
MUC2_HUMAN
ID MUC2_HUMAN STANDARD; PRT; 5179 AA.
AC Q02817; Q14878;
DT 01-JUN-1994 (Rel. 29, Created)
DT 01-NOV-1997 (Rel. 35, Last sequence update)
DT 20-AUG-2001 (Rel. 40, Last annotation update)
DE MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN MUC2 OR SMUC.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Intestine;
RX MEDLINE=94132002; PubMed=8300571;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT Identification of the amino terminus and overall sequence similarity
RT to prepro-von Willebrand factor.";
RL J. Biol. Chem. 269:2440-2446(1994).
RN [2]
RP SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC TISSUE=Colon;
RX MEDLINE=93016075; PubMed=1400449;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA Kim Y.S.;
RT "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT both upstream and downstream of its central repetitive region.";
RL J. Biol. Chem. 267:21375-21383(1992).
RN [3]
RP SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX MEDLINE=91358717; PubMed=1885763;
RA Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA Petersen G.M., Kim Y.S.;
RT "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT and polymorphism.";
RL J. Clin. Invest. 88:1005-1013(1991).
CC -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC AGENTS AT MUCOSAL SURFACES.
CC -!- SUBUNIT: MULTIMERIC.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC BRONCHUS, CERVIX AND GALL BLADDER.
CC -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC VARIES AMONG DIFFERENT ALLELES.
CC -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC OF SILKWORM HEMOCYTIN.
CC -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC ---

```
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;
```

```
Query Match             16.3%;  Score 1194;  DB 1;  Length 5179;
Best Local Similarity   30.6%;  Pred. No. 5.6e-40;
Matches  400;  Conservative   74;  Mismatches  515;  Indels  320;  Gaps   51;

Qy       4 KTLPIYLLLLLSVFVIQQVSSQDLSSC--AGRCG-----EGYSRDATCNCDYNCQHYMEC 56
           |  |||                :||  |  | :||       |   |:   :  |:  : |
Db    1169 KDRPIY-------------EEDLKKCVTADKCGCYVEDTHYPPGASVPTEETCKSCV-C 1213

Qy      57 CPDFKRVCTAELSCKGRCFESFERG-----RECDCDAQCKKYDKCC-----PDYESFCAE 106
             : ||  |   :|:     : |       |  :   :|:   |     |   :
Db    1214 TNSSQVVCRPE---EGKILNQTQDGAFCYWEICGPNGTVEKHFNICSITTRPSTLTTFTT 1270

Qy     107 VHNPTSPPS-SKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRT 165
           :  ||:| | :      :  | |: |||   ||          : |: | |
Db    1271 ITLPTTPTSFTTTTTTTTPTSSTVLSTT---PK-------LCCLWSDWINE--------- 1311

Qy     166 KKKPTPKPPVVDEAGSGLDNGDFK---------------VTTPDTSTTQH-NKVSTSPK 208
                 |   || |:|| :                  |  |  |  ||  ||
Db    1312 ----------DHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVS 1360

Qy     209 ---------------------------------ITTAKPINPRPSLPPNSDTSKETSL 233
                                             |||  |    || || : |:   :
Db    1361 VGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTLPPTT 1420

Qy     234 TVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAE 293
           | :  ||     ||||   |:|     ||:                ||:     |:|
Db    1421 TPSPPTT-----TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPIS 1459

Qy     294 TTTKGPALTTPKEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------AP 340
           |||  |  |||  ||      |||   | :||   | |||  | | |||  |        |
Db    1460 TTTTPPPTTTPSPPTTTPSPPTTTPSPPTTTTTTPPPTTTPSPPMTTPITPPASTTTLPP 1519

Qy     341 TTTKSAP-TTPKEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPK 390
           |||  |  | ||   | ||||  |  |||  |          ||||  | ||||  :  |  |||
Db    1520 TTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPS 1579

Qy     391 EPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTA 443
           |  |||  |  ||    |  |||  || ||   |  ||  |  |||  |           ||
Db    1580 PPTTTTPSPPTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTT 1639

Qy     444 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 503
             | |||   | |||   |  |||  |  ||   | |||| |:| ||    |||| : |
Db    1640 TPSPPPTTTTTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTPSPPTTTMTTP 1699

Qy     504 ---TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPT 560
             |||  |  |||  |: ||   | |||   |:|||   |  ||     ||||  |  |
Db    1700 SPTTTPSSPI-TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTT 1758
```

```
Qy     561 APKEPAPTTP-----KETAPTTP---------------------------------- 578
           |  |: | |       | ||||
Db    1759 TPLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWA 1818

Qy     579 ------KKLTP---------------------------------------------- 583
                 : |
Db    1819 ANISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCEC 1878

Qy     584 -TTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTT 641
            | |  :  || | | | |   :  ||   |||| |  :   |||          | ||
Db    1879 VTQPTTMTTTTTENPTPPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTG 1938

Qy     642 PKEP-----APTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTT 695
           : |       ||   | || |  :|  |||  |  |    | ||  : |    :  |||
Db    1939 TQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTT 1998

Qy     696 KEPTSTTSDKPAP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKP 754
             || | :    | |||  |  |    | ||  : |  |||  |  |    | ||  : |
Db    1999 VTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTP 2057

Qy     755 APKELAPTTTKGPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSE 813
             :  |||  || | :    | |||  |  |    | ||  : |  |||  |   | :
Db    2058 TTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTP 2116

Qy     814 VSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPA 862
             ||| |: |||   :|  :|  ::  |||   :     |             || |:||
Db    2117 TPTPTGTQTPTT---TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT 2173

Qy     863 ATKPEMTTTAKDKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATT 914
           |    |||     |    :   |||  |||  |  |    :| ||| ' || : :| |
Db    2174 TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTP 2233

Qy     915 TQVTSTTTQDTTPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKA 973
           |  | | |  |||  |||  |||: |  | |  :| ||| |       |  |    | :
Db    2234 TP-TGTQTPTTTP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT 2288

Qy     974 TTPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAM 1029
           |||     |  |   || |: | | |       ||||    :  |   |   |: :
Db    2289 TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTP 2348

Qy    1030 LQTTTR-PNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 1077
           | |: |  || :    | |     |    :||     |      |||
Db    2349 TPTGTQTPTTTPITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392


RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL        1     28       POTENTIAL.
FT   CHAIN        29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN       36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT       36    191       1.
FT   REPEAT      207    363       2.
FT   REPEAT      409    565       3.
FT   REPEAT      607    763       4.
FT   DOMAIN      771   1377       APPROXIMATE TANDEM REPEATS OF
FT                                T-P-S-D-E-P.
FT   DOMAIN     1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN     1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN     1495   1565       SLH 2.
FT   DOMAIN     1566   1625       SLH 3.
FT   DOMAIN     1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;


  Query Match             13.0%;  Score 950;  DB 1;  Length 1664;
  Best Local Similarity   31.0%;  Pred. No. 7.9e-31;
  Matches  312;  Conservative  104;  Mismatches  361;  Indels  230;  Gaps   55;

Qy     284 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 342
           |: :| | || :    |  |   ||||:          ||||:    ||  || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     343 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 398
             :  ||:|| || |    ||  ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     399 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 457
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     458 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 510
           :    |:  || : ||:|: ||:|| || | |  || : || |:   |    ||   ||:
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy     511 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 565
               || : || |: ||:|| ||      ||   || |:    |:   |    ||   ||
Db     980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy     566 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 618
            |:   || |    ||:: ||      ||  ::| |:   || :  ||  :||||: ||
Db    1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy     619 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 675
           ||| ||   :  |    || |:    |: ||   || |: ||:|  ||      ||
Db    1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy     676 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 734
           || |:   :| | :  || : ||| | | ::| ||      ||   || |:   || |:
Db    1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy     735 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 790
             ||   || |: ||::|            || | ||:| |:   || :  ||   |
Db    1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy     791 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 849
           | |:   :| |: ||| | || :    || : |||      || ||  | ||||
Db    1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305
```

```
Qy    850 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 907
Db   1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy    908 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 967
Db   1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy    968 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 1012
Db   1397 SGGSGGSGGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy   1013 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 1064
Db   1454 YLRGYPDGSFRPERNITRAEAAVIF-----------AKLL----GADESYGAQSASPYSD 1498

Qy   1065 LLRPH---------------------VFMPE--------VTPDMDYLPRVPNQGI----- 1090
Db   1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy   1091 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1137
Db   1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT--------------FKPQN 1600

Qy   1138 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1182
Db   1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

```
RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;
```

```
  Query Match           10.9%;  Score 795.5;  DB 1;  Length 1367;
  Best Local Similarity  27.4%;  Pred. No. 7.6e-25;
  Matches  318;  Conservative  111;  Mismatches  525;  Indels  207;  Gaps   51;

Qy     85 DCDAQCKKYDK---CCPDYESFCAEVHNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPN 141
Db    174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSES--STTTSS----- 226

Qy    142 KKKTKKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHN 201
Db    227 --------TSESSTTTSSTSESST-------------------------TTSSTSESSTS 253

Qy    202 KVSTSPKI-TTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEK 260
Db    254 SSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPCTKKKTTTSK-TCT---KK 309

Qy    261 TTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPALTTPKE------PTPTTP 312
Db    310 TTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPVTSSTTESSSAPVPTPSSS 362

Qy    313 KEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKE 368
Db    363 TTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVTSS----TTESS 418

Qy    369 PAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTP---TTPKEPA 425
Db    419 SAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSA 471

Qy    426 PTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEPAPTTTKEPSPTTPKEPAP 481
Db    472 PVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSSAPVPTPS-SSTTESSSAP 526

Qy    482 TTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP----TTPKEPAPT-TPKK 532
Db    527 APTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTPVTSSTTESSSAPVPTPSS 586

Qy    533 PAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPT 592
Db    587 STTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSAPVT----SSTTESSSAP- 637

Qy    593 TPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKAAAPNTPKEPAPTTPKEPA 646
Db    638 VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVT----SSTTESSSA 693

Qy    647 PTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDK 705
```

```
          | |    : ||   :||  ||  :   :   | ||          ||  |   :|||
Db    694 PVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPS-SSTTESS 748

Qy    706 PAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTLKEPAPTTPKKPAPKELAP 761
          || |   |  ::   | |   ||    ||  ||  :   :   | ||         :||
Db    749 SAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESSVAP 808

Qy    762 TTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKKPAPTTPETPPPTTSEVST 816
            |   :|  :      :||  ::  :   | |     ||    || :  |   : : | |
Db    809 VPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESSSAPVSSSTTESSVAPVPT 868

Qy    817 PT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTT 871
          |:     |:  |::|   |  || |  :  |     :| | ||  |  : ::|
Db    869 PSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYPGSQTETSVSSTTETTIVP 924

Qy    872 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT--TTQVTSTTTQDTTPFK 929
          |  |:     ||  |||     |     ||   :  :| |  ||  || ||  ||
Db    925 TKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTTVPTTTTTSVTTSSTTTIT 984

Qy    930 ITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDRATNSKATTPKPQKP---- 981
          |   | | :   ||   : ||||||      | |||        |::||  |  |
Db    985 TTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--------SESTTTSPTTPVTTV 1036

Qy    982 ------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEA 1028
                      ||   : |:|   | :|        |||    :|:    ||: |
Db   1037 VSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP--SVTTVTNFTPTT-ITTT 1093

Qy   1029 MLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQ 1088
          :  | |             |   |   |   :|    :       | | :   |
Db   1094 VCSTGT------------NSAGETTSGCSPKTVTTTVPCSTGTGEYTTEATTLVTTAVT 1140

Qy   1089 GIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGI 1148
            ::    |  ||   ||   | ||    |       |   |:| :| :||       :
Db   1141 TTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTLAPSAPVTPATN-----AV 1188

Qy   1149 PSPIDTVFTRC----NCEGKT 1165
          |: | |  | |    |  |:|
Db   1189 PTTITT--TECSAATNAAGET 1207
```

```
RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
```

```
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL       1     23       POTENTIAL.
FT   CHAIN       24   1255       MUCIN 1.
FT   DOMAIN      24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM  1163   1186       POTENTIAL.
FT   DOMAIN    1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN      81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN    1034   1151       SEA.
FT   CARBOHYD   957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC    19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC    20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC    20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC   126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                               ISOFORM).
FT   VARSPLIC  1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                               ISOFORM).
FT   VARSPLIC  1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT     2      2       T -> A (IN REF. 11).
FT   CONFLICT   134    134       P -> Q (IN REF. 9).
FT   CONFLICT   154    154       P -> Q (IN REF. 9).
FT   CONFLICT  1021   1021       S -> T (IN REF. 3).
FT   CONFLICT  1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE  1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;


  Query Match             8.9%;  Score 651;  DB 1;  Length 1255;
  Best Local Similarity   27.8%;  Pred. No. 3.2e-19;
  Matches  295;  Conservative   72;  Mismatches  460;  Indels  234;  Gaps   51;

Qy    231 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 289
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    290 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTI---K 329
          |    :||:|    |   ||                | |    ||||     |:    |
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy    330 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 383
          || :   ||   | ||| |   ||| :|  |       || |   ||| :|  ||   |
Db    126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy    384 SAPTTPKEPAPTTP----------KKPAPTTPKEP------APTTPKEPTPTTP------ 421
          ||| |  |  | |             :||| :   |       || |   | | |
Db    180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy    422 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 462
              ||| :  |        || |   |  |||        || |   |  | |
Db    240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy    463 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 517
          || |   |  | |      |:   |: || :|  ||   | ||| |   |:| :|  |
Db    300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy    518 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 565
              || |   |  | |       || |   |  | |       || |   |  |||
Db    357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy    566 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 625
          |:  :| |       |      || |   |  | |  |      | |   |  |
Db    417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy    626 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 677
          | |    :||:|   |  | |       || |   |  | |    |     :|| |   |
Db    473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy    678 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 725
          | |           :|||   ||      ||    :|||  | |     |:    ||
Db    529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy    726 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 771
          |   |  | |       :|| |   |  | |           :|||   ||    | ||
Db    584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy    772 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 815
            :|||  | |       :|| |   |  | |           :||| :  | ||    |
Db    643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700

Qy    816 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 868
          | |   | :       :  :||      | |    : ::|     |   | :| |
Db    701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753

Qy    869 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 921
          |   |:  | |  | :|       |:||    :|      |       |:   |
Db    754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807
```

```
Qy     922 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 977
Db     808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867

Qy     978 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 1037
Db     868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA----------PG 909

Qy    1038 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 1096
Db     910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy    1097 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1137
Db     956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1       1
FT   DOMAIN        67     870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                   P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;
```

```
  Query Match             7.7%;  Score 566.5;  DB 1;  Length 875;
  Best Local Similarity   27.2%;  Pred. No. 4.9e-16;
  Matches  274;  Conservative  103;  Mismatches  404;  Indels  227;  Gaps   54;

Qy     108 HNPTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKK 167
Db       3 HEPVYKP-------------VKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYGSKTNY 48

Qy     168 KPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVST-----SPKITTAKPINPRPSLP 222
Db      49 LPLAK---------------KLSSYKPIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYP 92

Qy     223 PNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTS 282
Db      93 P----------TYKSKPTYKPKIT----------YPPTYKAKPSYPSSYKPKKTYPPTY 131

Qy     283 KVLAKPTPKAETTTKGPALTTPKEPTPTT----PKEPASTTPKEPTPTTIKSAPTTP--- 335
Db     132 K------PKLTY----PPTYKPKPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSYPPTY 181

Qy     336 --KEPAPTTTKSAPTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEP 392
Db     182 KAKPSYPPTYKAKPSYP----PTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTYKAKP 237

Qy     393 A--PTTPKKPA--PTTPKEPA--PTTPKEPT-PTTPKEPAPTTKEPAPTTPKEPAPTAPK 445
Db     238 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYP----PTYKA 293

Qy     446 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 495
Db     294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy     496 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 551
Db     354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy     552 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 611
Db     410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy     612 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 665
Db     458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy     666 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 721
Db     507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy     722 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 773
Db     552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy     774 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 822
Db     612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy     823 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 871
Db     668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy     872 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 931
Db     727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy     932 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 988
Db     778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826
```

```
Qy    989 TSTKKPKTMPRVRKPKTTPTP-----------RKMTSTMPELNPTSR 1024
          :: |   : |   | ||: |           :  ||  |   ||:|
Db    827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874


RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;


  Query Match             7.6%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity   32.6%;  Pred. No. 3.3e-15;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy    214 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 273
          |:   | ||| |  |     |   |||  |:  | : |             | | ||||
Db    445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy    274 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 328
          |     ||      |  |    |: | ::| |   || | |  :   ::||     || :
Db    488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy    329 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 387
            || |  ||     ||   |     |:|  |||| |   |  :||  ||   | ||
Db    540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy    388 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 447
            ::|:  | |   ||   |    : ::||    |   || |   || | | :  :|:|
Db    597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy    448 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 507
           |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db    648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy    508 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 562
          |    :|:|  |||| |    : :||:     || | |   |: :|   :| |   |
Db    702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy    563 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 612
          ::|  :| | | ||     :|:|||  | |||      || | :    : |:|  | | |
Db    758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy    613 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 672
            :||:|  :| |           || | |  || | |       ||  :| | | ||
Db    815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855
```

```
Qy    673 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 726
Db    856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy    727 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 784
Db    912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy    785 PTTPKEPAPT 794
Db    967 PASCKSPRPS 976


RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73        H-A-P-P.
FT   REPEAT      148    151        H-A-P-P.
FT   DOMAIN      229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235        1.
FT   REPEAT      236    242        2.
FT   DOMAIN      205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             7.5%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 1.5e-15;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy    292 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 340
Db     24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy    341 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 391
Db     81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy    392 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 446
Db    141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy    447 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 501
Db    194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy    502 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 555
Db    250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy    556 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 605
Db    310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy    606 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 662
Db    370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy    663 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 722
Db    428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy    723 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 782
Db    482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy    783 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 840
Db    539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy    841 PTPKALENSPKEPGVPTTKTPAATKP 866
Db    599 PSP------------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
```

```
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT.entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;
```

```
Query Match             7.3%;  Score 533;  DB 1;  Length 1087;
Best Local Similarity   27.5%;  Pred. No. 1.2e-14;
Matches  216;  Conservative   81;  Mismatches  330;  Indels  158;  Gaps   38;
```

```
Qy    177 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 231
Db    404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy    232 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 291
Db    450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy    292 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 351
Db    510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy    352 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 401
Db    565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy    402 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 446
Db    620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy    447 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 498
Db    677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy    499 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 553
Db    730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy    554 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 613
Db    790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy    614 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 661
Db    849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy    662 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 719
Db    905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964

Qy    720 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 779
Db    965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP------------ 1006

Qy    780 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 839
Db   1007 --------KKEEMPAAPEK-----------------KDTKEEKTTESRKPEEK------ 1034

Qy    840 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 899
Db   1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy    900 TTTEK 904
Db   1083 TKGEK 1087
```

```
RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
```

```
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;


  Query Match             7.2%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 1.6e-14;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy     308 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 366
           ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db     599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy     367 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 413
           :  |  |   |  ::  | |:|| :  | :||  |       ||:|| : :    |:||
Db     652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy     414 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 472
           : :   ||  || ::    |:|| : :       | :||  || ::      |     :
Db     712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy     473 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 528
           ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db     765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy     529 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 583
           ||    | :||  |  ::      |      | : |  || ::   | |:||       ||
Db     818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy     584 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 633
           :|| :  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db     877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy     634 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 693
           |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db     936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

Qy     694 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 752
               || :|  | :|     |:|| :  | :||  || ::    |  | :   | :||
Db     984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy     753 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 808
           |     | ::  |  || :|  | :||      :| :||  || ::    |  || ||
Db    1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy     809 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 867
           :::  |||:|  : :: | :|  ||| : ||  ||               |||  :
Db    1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy     868 MT 869
           |
Db    1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
```

```
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;


  Query Match             7.2%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 1.3e-14;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy    279 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 334
          :| |  :| | ||    |      ||   |       |  | : ||  | ||     | |
Db      8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy    335 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 389
          |||     :|  ||     ||     | |      || :    |              |||  |
Db     62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy    390 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 435
          |  ||    |       | |   :  ||   ||       : |  || ||   |
Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    436 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 487
          ||  |      ||      ||   : |||    | |  | |:  ||     :   :
Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    488 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 543
          |    ||   ||  |::    |   ||  : .||  |    |   : : |         ||
Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    544 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 584
           ||         : |  |     ||   ||   ||      ||        | |||  : |
Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    585 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 640
          | : ||     | | |  || | |:      || |   ||:   :|   || : ||
Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    641 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 689
          |    || |    | | ||   :|| |:             ||:     ||
Db    414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    690 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 741
           ||  :  | ||:  |  | |    || |       | |  ||  |    ||    |
Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    742 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 787
          | |      |  |  | |  |  ||:    :   ||:|  | | |    | |   :
Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    788 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 837
          || |  :| |     | |: :  : |   : |  | :: |           | ||
```

```
Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    838 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 895
          : |  |:|:     :| :  |: |  |:         |      :: ||| |  ||| :::
Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    896 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 948
                :    :|       | |         | | | :    : :  | | ::
Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    949 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 1008
                : |: |:        : |  |  : ||   : |  :  |     |   | ||
Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy   1009 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 1051
          | : :|   |   :   :|:   |:     ||    :| |:      |||
Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864


RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN       21     41       NONREPETIVE LINKER.
FT   DOMAIN      124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                                P-[PST]-[ST].
FT   DOMAIN      184    192       NONAPEPTIDE 1.
FT   DOMAIN      213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;
```

```
  Query Match             7.1%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity   29.2%;  Pred. No. 3.7e-14;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy    175 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 225
          |   | ||   | :| | | :    ::|       ||: | | |  :|   | |  |
Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy    226 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 285
            :|    |   :|           | :|    : |  | : ::    | |      |:
Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    286 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 344
            || |       | :| |  ||    |  | |  ||   | | |     ||   | | |
Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    345 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 399
          |: || :|  |   | :|  | |    ||  ::|: |     |  ||   | |  ||
Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    400 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 452
            |:  |  |        |:  | |  ||   | | :| |: |    ||  ||   | |
Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    453 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 507
          | | |: |    |  |:   ||   ||  |  |:       :|   ||: ||  | |
Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    508 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 557
            |  |  |      |  |    |||  :||  ||        :| |: |    | ||  |
Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    558 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 612
            ||   | | |  |       : |   ||   |::     || |: |    |
Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    613 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 663
          ||    |  |    ||   | | | :|:  || :  |: |     |   ||:   |:| |
Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    664 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 720
            |  || | | |  |      ||    |    | ||  ||     :  ||  || : | | |
Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    721 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 776
                      || | | | | | |: |    :||:   :  : :  ||:  | |
Db    584 --------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    777 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 825
          | |  || | | | | | |: |   :|  | | |         |    |: | | : :   |
Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    826 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 881
            |    |: :  || |     ||    || |    |||    : | ||    | :
Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    882 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 941
            : | |  | |     :   | |   |  ::   ::    |    |   :: |     ||
Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    942 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 996
          :          :   | |||     | |:: ||:    ||  ||    |: ||| |
Db    786 I--------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy    997 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 1038
            |  ||| :  || :   :  |   |      |     : |:|
Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871


RESULT  12
MUC1_XENLA
```

```
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
```

```
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;
```

```
  Query Match             7.0%;  Score 509.5;  DB 1;  Length 662;
  Best Local Similarity   28.1%;  Pred. No. 6.6e-14;
  Matches  223;  Conservative   48;  Mismatches  276;  Indels  247;  Gaps   30;

Qy    248 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 307
          ||    :  ||: | :|: : : |||:|       :| | ||  |     |   ||
Db      3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy    308 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 367
              |: |    |      ||||   || |    ||||    ||||    ||||
Db     52 --------AAATAAAETTAAAGEAPTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy    368 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 425
          || |   | |||  | ||||      :|    ||||    | :  :  : :: :|
Db    104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy    426 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 470
            |: :|   :   | |:      ||     ||          :|   : :|   ||||
Db    164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy    471 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 530
              |||| : | |||    ||||           ||| :  |||
Db    219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy    531 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 590
                      ||  :  ||||            || | |  ||    ||||
Db    246 ------------TTTTKATPTTT------------TTTKATTTTT----TPTT----- 270

Qy    591 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 650
                |||    | |||   | |||
Db    271 ---------TTTTTKATTTPTTTTTTTP------------------------------- 289

Qy    651 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 710
                  |  ||   |  ||       | |          :|    :    :
Db    290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy    711 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 760
```

```
          ||     : | |              ||:     |    |    |    :|    : :
Db    329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy    761 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 820
          : ||   :||   |  |   | |||   |  ||  |   |||      |:  :|||||
Db    385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTPTTT 436

Qy    821 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 880
            ||                | ||           || ||  | |  |||   | |
Db    437 TTTTT---------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy    881 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 940
            ||| |||  |     | |||: |       |||  |:|||  ||  | ||   ||
Db    464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy    941 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 996
            |||| | |||    | |    |  ||       : |  :|    :|  : |    :
Db    511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy    997 MPRVRKPKTTPTPR 1010
          :|:|   |  |: |
Db    567 LPQVADCKVAPSSR 580
```

```
RESULT  13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
```

```
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;


  Query Match             6.9%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 2.8e-13;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     307 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 360
           |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     361 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 418
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     419 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 478
           |:|   :||:   :||:|   :||:|   :||:|   :||:|            |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     479 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 538
            :||:   :||:    :||:   :||:|   |||:    :||:|  :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     539 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 598
              :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db    1726 SY-SPTSPSY-SPT$PSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     599 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 656
           ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     657 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 715
           :|   :||:|| |          ||:|   |:  |  ||: |   | || | :|||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy     716 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 774
           ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     775 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 832
            :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     833 STPE 836
             |  |
Db    1966 SDEE 1969


RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;


  Query Match             6.8%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 4.3e-13;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;
```

```
Qy    307 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 360
          |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db   1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy    361 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 418
          | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :|||:|   :|
Db   1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy    419 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 478
          |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    479 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 538
           :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    539 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 598
             :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    599 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 656
          ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    657 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 715
          :|   :||:|| |            ||:|   |:  |  | : |    | || | :||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy    716 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 774
          ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    775 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 832
           :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    833 STPE 836
          |  |
Db   1966 SDEE 1969
```

```
RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER      1      1
FT   DOMAIN     276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT   164    164       L -> I (IN REF. 2).
FT   CONFLICT   185    185       I -> S (IN REF. 2).
FT   CONFLICT   193    193       L -> T (IN REF. 2).
FT   CONFLICT   199    199       M -> T (IN REF. 2).
FT   CONFLICT   346    346       K -> N (IN REF. 1).
FT   CONFLICT   373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT   482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT   485    485       P -> S (IN REF. 2).
FT   CONFLICT   570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT   591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT   727    727       A -> V (IN REF. 4).
FT   CONFLICT   757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT   769    769       T -> L (IN REF. 2).
FT   CONFLICT   775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;


 Query Match             6.8%;  Score 498;  DB 1;  Length 831;
```

```
  Best Local Similarity   27.2%;  Pred. No. 2.3e-13;
  Matches  205;  Conservative   63;  Mismatches  320;  Indels  166;  Gaps   38;

Qy    146 KKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVST 205
          :|:::| ||          |    | |           |  |  |:  |  ||  | || :
Db    170 RKLLEGEEC---------RIGFGPIP---------FSLTEGLPKI--PSMST--HIKVKS 207

Qy    206 SPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 265
            ||   :             : ||| :   :   ::  |  |  :     |:|:  : :
Db    208 EEKIKVVE------------KSEKETVIVEEQTEEIQVTEEVTEEEDKEAQGEEEEEAEE 255

Qy    266 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----TPTTPKEPA-STTP 320
            :    ||     | ::  | |    |  || |     | |    :|   | || : :|
Db    256 GGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKSPAEAKSPAEAKSP 306

Qy    321 KEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTKEP----APTTPKE 368
          |     :  :|   | ||    |   || |   | ||    | | | |    :|   |
Db    307 AEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVKSPGEAKSPAEAKS 365

Qy    369 PA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA----PTTPKEPA--- 409
          ||    |   | ||    | : || |   | ||    |   | ||    |   | ||
Db    366 PAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATVKSPVEAKSPAEVK 424

Qy    410 -PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTAPKKPAPTTPKEPA 456
           | | | |    :|   | || : | |    :|   | ||    |   | ||    | ||
Db    425 SPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVVAKSPAEA--KSPA 481

Qy    457 PTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEPAPTTTKS-----A 502
            |   | :  :. ||   | ||    |   ||     :|   | |     ||     :
Db    482 GAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSPVKEGAKSLAEAKS 541

Qy    503 PTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPA 558
          |   | |     | ||    |   | |     | |      :      | ||    |:||
Db    542 PEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPEAKPPAKEEAKRPA 601

Qy    559 PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTP 618
              |   :  || | :  |: | |  ||:||   |  :|   ||:     :||
Db    602 DIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV---KAKEPPKKVE 654

Qy    619 EEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPA 678
          ||  | |||     : |: ||   ::|     |||    ||::     |  |    ||
Db    655 EEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEAKKEEA----KEKK 709

Qy    679 PTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGT 738
             |::  | :|      :|     |  |            |||:  : ||     ||
Db    710 AAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKPSEKE----KPK-- 751

Qy    739 APTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKK 798
             ||  |  |:|   ||           ||  |     ||  |    ::       |
Db    752 -----KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAKAKEED-------K 790

Qy    799 PAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 832
            |  |  |    :| |: |  |:     |:|::
Db    791 GLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824


Search completed: April 26, 2002, 16:33:14
Job time: 624 sec
```

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:26:07 ; Search time 61.21 Seconds
(without alignments)
500.359 Million cell updates/sec

Title: AA7
Perfect score: 7323
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1361

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	7.1	805	4	US-09-103-429A-4	Sequence 4, Appli
2	510.5	7.0	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	6.9	744	6	5202236-25	Patent No. 5202236
4	498.5	6.8	786	4	US-09-103-429A-3	Sequence 3, Appli
5	496.5	6.8	1721	3	US-08-700-651-5	Sequence 5, Appli
6	496.5	6.8	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	6.7	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	6.7	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	6.5	652	6	5202236-13	Patent No. 5202236
10	455.5	6.2	960	4	US-09-219-849-5	Sequence 5, Appli
11	428	5.8	1185	4	US-09-041-886-23	Sequence 23, Appl
12	424.5	5.8	1867	2	US-08-479-537A-5	Sequence 5, Appli
13	424.5	5.8	1867	4	US-09-083-116-5	Sequence 5, Appli
14	424.5	5.8	2035	2	US-08-479-537A-2	Sequence 2, Appli
15	424.5	5.8	2035	4	US-09-083-116-2	Sequence 2, Appli
16	420.5	5.7	907	3	US-08-783-774-2	Sequence 2, Appli
17	420.5	5.7	907	5	PCT-US95-04611A-19	Sequence 19, Appl
18	419.5	5.7	2476	2	US-08-276-967-2	Sequence 2, Appli
19	417	5.7	829	1	US-08-642-255-132	Sequence 132, App
20	417	5.7	829	1	US-08-397-633A-53	Sequence 53, Appl
21	417	5.7	837	1	US-08-175-155-68	Sequence 68, Appl
22	417	5.7	837	1	US-08-477-509B-103	Sequence 103, App
23	417	5.7	837	1	US-08-642-255-101	Sequence 101, App
24	417	5.7	837	2	US-08-707-237A-75	Sequence 75, Appl
25	417	5.7	837	3	US-08-482-085B-103	Sequence 103, App
26	417	5.7	897	1	US-08-397-633A-50	Sequence 50, Appl
27	404.5	5.5	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	5.5	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	5.5	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	5.4	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	5.4	682	1	US-08-397-633A-36	Sequence 36, Appl
32	385	5.3	1231	3	US-08-904-263A-4	Sequence 4, Appli
33	381.5	5.2	1537	1	US-08-325-267A-2	Sequence 2, Appli
34	381	5.2	1848	4	US-08-296-791-6	Sequence 6, Appli
35	381	5.2	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
36	379.5	5.2	1187	1	US-08-320-559-28	Sequence 28, Appl
37	379.5	5.2	1187	3	US-08-545-860D-28	Sequence 28, Appl
38	379.5	5.2	1187	5	PCT-US94-04496-28	Sequence 28, Appl
39	379.5	5.2	1210	1	US-08-320-559-26	Sequence 26, Appl
40	379.5	5.2	1210	3	US-08-545-860D-26	Sequence 26, Appl
41	379.5	5.2	1210	5	PCT-US94-04496-26	Sequence 26, Appl
42	376	5.1	960	4	US-09-219-849-6	Sequence 6, Appli
43	371	5.1	761	2	US-08-707-237A-84	Sequence 84, Appl
44	371	5.1	762	1	US-08-642-255-114	Sequence 114, App
45	371	5.1	762	1	US-08-397-633A-26	Sequence 26, Appl

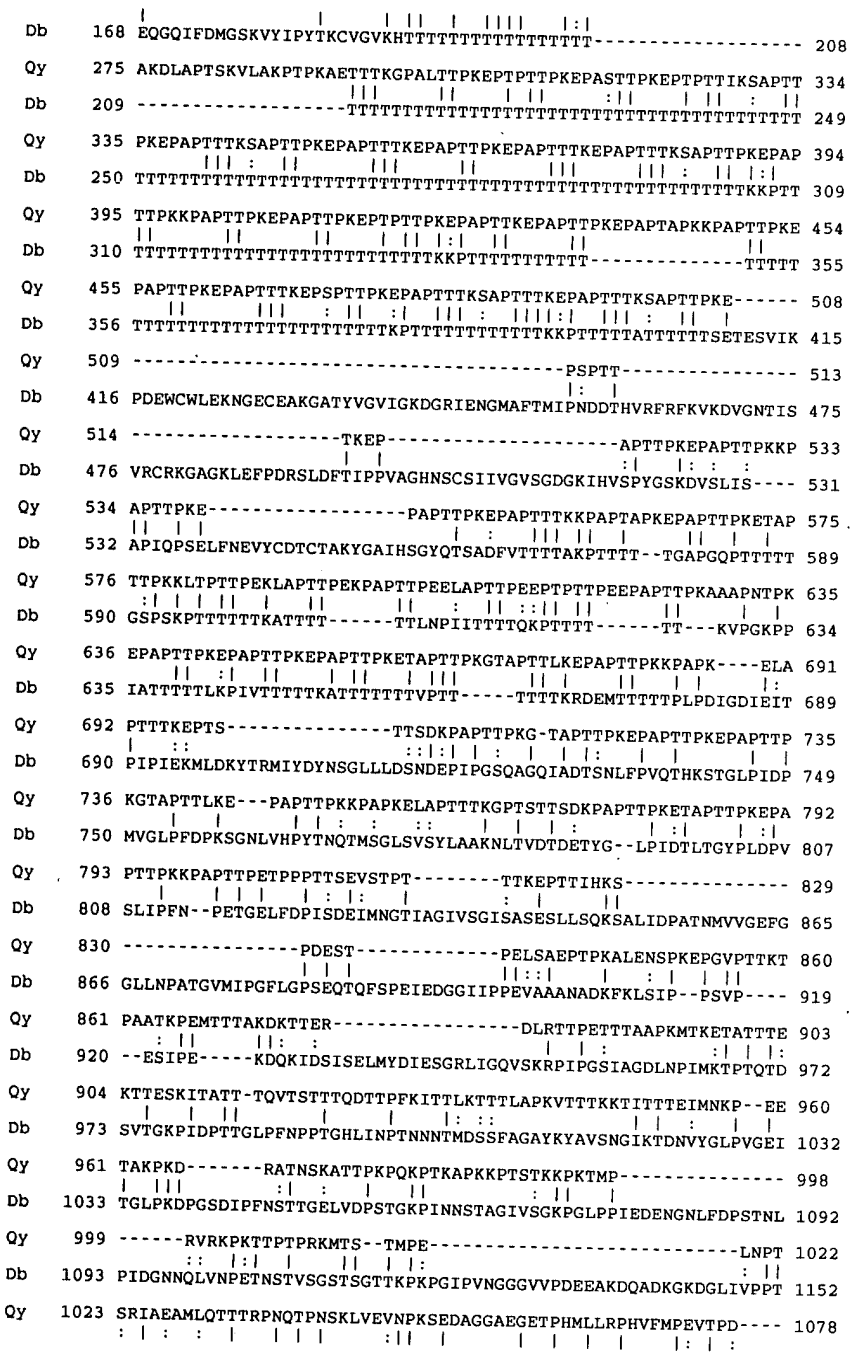
ALIGNMENTS

```
RESULT   1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-4


  Query Match          7.1%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity  27.8%;  Pred. No. 3e-25;
```

```
  Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy    281 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 335
          |: |    |:     | ||| |    |  | |     | |   |              |
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    336 KEPAPTT------------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 383
          :: || |             :  | |  || ||    || | :  |||||:  |||||
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    384 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 443
           ||||  :   ||            ||||  : || |     | ||: |  ||
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    444 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 503
              ||||  : |||| : |||||:         |||||: |||||:  ||||   |
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    504 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 547
          |||   :|  |  || |||    || |||  ||||:    ||  |
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    548 ----------------------------------------------------EPAPTTTK 555
                                                             |||  |:
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    556 K----------------PA---------------------------------------- 558
                             ||
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    559 --------------------------------PTAP--KEPAPT-----------TPKET 573
                                           || |   || |             | |:
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    574 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 599
                                  :| |    :| |           :   |     ||
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    600 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 659
          |     |||       |:|   || | |  |||  |    ||    ||||     ||||
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    660 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 716
            :|||   |   |||     ||||    |  |:  | |  ||:  :   |||     ||
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    717 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 775
          ||     |||      |:|   ||||   ||| |           | |  ||:  |:  |
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT-----------TVTVPPTAAPTTAAP 641

Qy    776 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 835
          ||     |  | |   ||||   ||| |  | ||| :  : |     ||  |        |
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    836 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 867
            :| |           | ||  || | ||
Db    686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;   INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             7.0%;  Score 510.5;  DB 3;  Length 1837;
  Best Local Similarity   20.8%;  Pred. No. 4.1e-24;
  Matches  314;  Conservative  109;  Mismatches  530;  Indels  559;  Gaps   52;

Qy    131 STTKRSPKPPNKKKT---------KKVIESEEITEVKDNKKNRTKKKPTPKPPVVDEAGS 181
          | |    | |||: |         |:    :  |  : :        | |  |     |:
Db    174 SITSGELKDPNKQATISGSRSCGWKQGYSIDSSTGFRVDSITGLPTDPYPNCPFNPVTGN 233

Qy    182 GLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKP--------INPRPSLPPNSDTSKE--- 230
            :    | | |:|    :  |  | :  | :|         ::|: : | ||: | |
Db    234 LVSRSTGK-TIPNT----YAGVYRSNETKTTEPSANTNFLLVDPKINAPCNSENSFEQVQ 288

Qy    231 -----TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 285
               : : :     |  | ||||   |:|    ||:   | :   |:      |:
Db    289 IFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 348

Qy    286 AKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKS 345
            |    |||      ||    | || |:| :||    | ||  :  ||       ||| :
Db    349 TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 408

Qy    346 APTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP 405
            ||      ||||:|  |       |||      ||| :  ||       || |||  ||
Db    409 TTTTTTTTTTTTTKKPTTT-------TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTT- 460

Qy    406 KEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAP 465
              ||    | ||      ||     ||    | |  |||  ||      || |:|
Db    461 -----TTTTTTTTTTTTTTTTTTTTTTTTTTTTATTTTKKPT-TTTTTTTTTTTKKPTT 514

Qy    466 TTTKEPSPTT---------------------------------------------------- 475
          |||   : ||
Db    515 TTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPND 574

Qy    476 ---------------------------------------------------------- 475

Db    575 DTHVRFRFKVKDVGNTISVRCGKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGGGKI 634

Qy    476 ---------------PKEPAP---------------------------TTTKSAPTTTK 492
                         | :|                             |||  : ||||
Db    635 HVSPYGSKDVSLISAPIQPCELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT 694
```

```
Qy    493 EPAP----TTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE 548
Db    695 TGAPGQPTTTTTGSPSKP--TTTTTTKATTTTTILNPIITTTTQKPTTTTTTKVPGKPPI 752

Qy    549 PAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE-ELAP 607
Db    753 ATTTTTLKPIVTTTTTKATTTTTTTVPT-----TTTTTKRDEMTTTTTPLPDIGDIEITP 807

Qy    608 TTPEEPTP--------------TTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTPK 651
Db    808 IPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDPM 867

Qy    652 EPAPTTPK--------------------------ETAPT--TPKGTAPTTLKEPAPTT 681
Db    868 VGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYGLPIDTLTGYPLDPVSLI 927

Qy    682 PKKPAPKELAPTTTKE--------------PTSTTSDKPAPTTPKGTA----------P 716
Db    928 PFNPETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSAPIDPATNMVVGEFGGLLNP 987

Qy    717 TT---------PKEPAPTTPK-EPAPTTP-----------KGTAPTTLKEPAP------- 748
Db    988 ATGVMIPGSLGPSEQTPFSPEIEDGGIIPPEVAAANADKFKLSIPPSVPESIPEKDQKID 1047

Qy    749 ---------------TTPKKPAPKELAPTTT---KGPTSTTS--DKPA-PTTPKETAPT 786
Db   1048 SISELMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTDSVTGKPIDPTTGLPFNPP 1107

Qy    787 T-----------------------------------PKEPAPTTPKKPAPTTPETPPPT 810
Db   1108 TGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVDEITGLPKDPVSDIPFN--ST 1165

Qy    811 TSEVSTPTTTK--------------------------EPTTIHKSPDESTPELSAEPTP 843
Db   1166 TGELVDPSTGKPINNYTAGIVSGKRGLPPIEDENGNLFDPST--KLPIDGNNQLVNPETN 1223

Qy    844 KALENS------PKEPGVPTT----------------------KTPAATKPEMTTTAK 873
Db   1224 STVSGSTSGSTKPK-PGIPVNGGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQY 1282

Qy    874 DKTTERDLRTTPETTTAAP-----KMTKETATTTEKTTESKITATTTQVTSTTTQDTTPF 928
Db   1283 SNTTGNII--NPETGKVIPGSLPGSLNYPSFNTPQQTDE--ITGKPVDTVTGLPYDPSTG 1338

Qy    929 KITTLKTTTLAP------KVTTTKKTITTTEIMNKP--EETAKPKD------RATNSKAT 974
Db   1339 EIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1398

Qy    975 TPKPQKP--------------TKAPKKPTSTKKP----------KTMPRVRKPKT----- 1005
Db   1399 DPSNKKPIPGSHSGFINGTSGEQSHEKDPSTGKPLDPNTGLPFDEDSGSLINPETGDKLQ 1458

Qy   1006 -------TPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPK-SEDA--- 1054
Db   1459 GSHSGTFMPVPGKPQGENGGIMTPEQILEAL-------NKLPTSNEVNISPRPSSDAVPD 1511

Qy   1055 ----------------------GGAEGETPHMLL------RPHVFMPEVTPDMDYLPRV 1085
Db   1512 RPTNTWWNKISGQTFQVDGKKTIPGSAASVIHTALGTPTQTDPTTGLP--SDPSTGLPFI 1569

Qy   1086 PNQGIIINPMLSD--------------ETNIC------------NGKPVDGLTTL---RN 1116
Db   1570 PGFNVLVDPQTGEQIKGSVPYVSLYVKEKNIVTEAAYGLPVDPKTGFPIDPISYLPFAKN 1629

Qy   1117 GTLV-AFRGHYF 1127
Db   1630 GELIDPISGKYF 1641


RESULT   3
5202236-25
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
;      LENGTH: 744
5202236-25


  Query Match             6.9%;  Score 506.5;  DB 6;  Length 744;
  Best Local Similarity   29.9%;  Pred. No. 2.6e-24;
  Matches  239;  Conservative   78;  Mismatches  308;  Indels  175;  Gaps   50;

Qy    207 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 266
Db     24 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 53

Qy    267 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 322
Db     54 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    323 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 379
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    380 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 431
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    432 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 480
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    481 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 533
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    534 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 585
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    586 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 634
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    635 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 688
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    689 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 747
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    748 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 802
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    803 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 849
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    850 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 895
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719
```

```
Qy    896 KETATTTEKTTESKITATTT 915
          :   |   | ::| |  :|
Db    720 YKAKPTYPSTYKAKPTYPST 739


RESULT   4
US-09-103-429A-3
; Sequence 3, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  786 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             6.8%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 8.8e-24;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    288 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 347
          | | || ||: || |      ||| : | :||           ||||  :   |||  ||
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    348 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 407
          || :  |||||:  | ||    ||||| : ||||| ||||  : |||| :  |||| :
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    408 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 464
           |||| : ||  |      |    || |||    | |   || |||  ||||: |  |
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    465 ----------------------------------------------PTTTKEPSPTTPK- 477
                                                         |   :  ||   :
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    478 ----EPAPTTT---------------------------------------------------- 484
              |||  |
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    485 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 517
                                      :|      | |  |    || | |  |     |
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    518 APTT-----PKEP----------------------APTTPKKPAPTT----------- 537
          |  :     | |                      :| |    :| |
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    538 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 594
               |||      |||      ||||   |:   |  | | |   : |||     ||
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    595 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 652
            ||||     ||||    :|||  | ||    ||| |  |: |   |    |||
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    653 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 712
          |||      ||||     |||:  |    |  ||  || ||                 | |
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    713 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 772
            ||||   ||| |     ||     || |:                 || ||  | :|||
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    773 DKPAPTTPKETAPTTPKEPAPTTPKKP 799
          || | |     |  | | |  |  |
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   5
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5


  Query Match             6.8%;  Score 496.5;  DB 3;  Length 1721;
  Best Local Similarity   22.1%;  Pred. No. 2.9e-23;
  Matches  295;  Conservative   86;  Mismatches  495;  Indels  459;  Gaps   43;

Qy    110 PTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKKKP 169
          || | |:    | :| :   :|| |  |                  |     : | ||
Db     82 PTDPYSNCPFNPVTG-NLVSRSTGKTIPN---------------TYAGVYRSNETK--- 121

Qy    170 TPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSK 229
          | :|      ||    | : | | |  :|   | :   |||          : | ||: |
Db    122 TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------NAPCNSENSF 167

Qy    230 E--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTS 274
```



```
Db    168 EQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTT------------------ 208

Qy    275 AKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 334
Db    209 ------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 249

Qy    335 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 394
Db    250 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 309

Qy    395 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKE 454
Db    310 TTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT-------------TTTTT 355

Qy    455 PAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKE------ 508
Db    356 TTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIK 415

Qy    509 ------------------------------------PSPTT---------------- 513
Db    416 PDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTIS 475

Qy    514 ------------------TKEP--------------------APTTPKEPAPTTPKKP 533
Db    476 VRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKDVSLIS---- 531

Qy    534 APTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAP 575
Db    532 APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGAPGQPTTTTT 589

Qy    576 TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPK 635
Db    590 GSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT---KVPGKPP 634

Qy    636 EPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK----ELA 691
Db    635 IATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPLPDIGDIEIT 689

Qy    692 PTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTTPKEPAPTTP 735
Db    690 PIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDP 749

Qy    736 KGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA 792
Db    750 MVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDTLTGYPLDPV 807

Qy    793 PTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS--------------- 829
Db    808 SLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPATNMVVGEFG 865

Qy    830 ----------------PDEST------------PELSAEPTPKALENSPKEPGVPTTKT 860
Db    866 GLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP--PSVP---- 919

Qy    861 PAATKPEMTTTAKDKTTER----------------DLRTTPETTTAAPKMTKETATTTE 903
Db    920 --ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTD 972

Qy    904 KTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKP--EE 960
Db    973 SVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVGEI 1032

Qy    961 TAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP--------------- 998
Db   1033 TGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENGNLFDPSTNL 1092

Qy    999 ------RVRKPKTTPTPRKMTS--TMPE--------------------------LNPT 1022
Db   1093 PIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPT 1152

Qy   1023 SRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD---- 1078
```

```
Db   1153 NSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGK 1206

Qy   1079 ----MDYLPRVPNQGIIINP-------------MLSDETNICNGK------------- 1106
Db   1207 PVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPR 1266

Qy   1107 -PVDGLTTLRNGTLV 1120
Db   1267 DPVSGLPQLPNGTLV 1281
```

```
RESULT   6
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6
```

```
  Query Match           6.8%;  Score 496.5;  DB 3;  Length 1721;
  Best Local Similarity  22.1%;  Pred. No. 2.9e-23;
  Matches  295;  Conservative   86;  Mismatches  495;  Indels  459;  Gaps   43;

Qy    110 PTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKKKP 169
Db     82 PTDPYSNCPFNPVTG-NLVSRSTGKTIPN---------------TYAGVYRSNETK--- 121

Qy    170 TPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSK 229
Db    122 TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------NAPCNSENSF 167

Qy    230 E---------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTS 274
Db    168 EQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTT------------------ 208
```

```
Qy    275 AKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 334
Db    209 ------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 249

Qy    335 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 394
Db    250 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 309

Qy    395 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKE 454
Db    310 TTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT-------------TTTTT 355

Qy    455 PAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKE------ 508
Db    356 TTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIK 415

Qy    509 ---------------------------------------PSPTT---------------- 513
Db    416 PDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTIS 475

Qy    514 ------------------TKEP--------------------APTTPKEPAPTTPKKP 533
Db    476 VRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKDVSLIS---- 531

Qy    534 APTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAP 575
Db    532 APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGAPGQPTTTTT 589

Qy    576 TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPK 635
Db    590 GSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT---KVPGKPP 634

Qy    636 EPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK----ELA 691
Db    635 IATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTPLPDIGDIEIT 689

Qy    692 PTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTTPKEPAPTTP 735
Db    690 PIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTHKSTGLPIDP 749

Qy    736 KGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA 792
Db    750 MVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDTLTGYPLDPV 807

Qy    793 PTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS--------------- 829
Db    808 SLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPATNMVVGEFG 865

Qy    830 ----------------PDEST-------------PELSAEPTPKALENSPKEPGVPTTKT 860
Db    866 GLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP--PSVP---- 919

Qy    861 PAATKPEMTTTAKDKTTER----------------DLRTTPETTTAAPKMTKETATTTE 903
Db    920 --ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNPIMKTPTQTD 972

Qy    904 KTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKP--EE 960
Db    973 SVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVGEI 1032

Qy    961 TAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP--------------- 998
Db   1033 TGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENGNLFDPSTNL 1092

Qy    999 ------RVRKPKTTPTPRKMTS--TMPE--------------------------LNPT 1022
Db   1093 PIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPT 1152

Qy   1023 SRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD---- 1078
Db   1153 NSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGK 1206
```

```
Qy   1079 ----MDYLPRVPNQGIIINP-------------MLSDETNICNGK-------------- 1106
Db   1207 PVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPR 1266

Qy   1107 -PVDGLTTLRNGTLV 1120
Db   1267 DPVSGLPQLPNGTLV 1281
```

RESULT 7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
; GENERAL INFORMATION:
; APPLICANT: Hoffman, Stephen L.
; APPLICANT: Charoenvit, Yupin
; APPLICANT: Hedstrom, Richard
; APPLICANT: Khusmith, Srisin
; APPLICANT: Rogers IV, William O.
; TITLE OF INVENTION: Protective malaria sporozoite surface protein
; TITLE OF INVENTION: immunogen and gene
; NUMBER OF SEQUENCES: 2
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: A. David Spevack
; STREET: NMRDC Building 1 T-12 National Naval
; STREET: Medical Center
; CITY: Bethesda
; STATE: MD
; COUNTRY: USA
; ZIP: 20814-5044
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.24
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/638,431
; FILING DATE: 19910110
; CLASSIFICATION: 424
; ATTORNEY/AGENT INFORMATION:
; NAME: Spevack, Avrom D.
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (301) 295-6759
; TELEFAX: (301) 295-4033
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 826 amino acids
; TYPE: AMINO ACID
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-07-638-431-2

Query Match 6.7%; Score 488.5; DB 1; Length 826;
Best Local Similarity 26.4%; Pred. No. 3.9e-23;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    324 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 376
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    377 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 432
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    433 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 490
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    491 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 548
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492
```

```
Qy    549 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 607
          |:         |  |  |  |: | |  |: |||                    |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    608 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 666
          : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    667 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 726
                    |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    727 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 786
          ||||                |  : ||:|    : |   : |     :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    787 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 840
          |  |             |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    841 PTPKALENSPK------------------EPGVPTTK 859
          |     |  ||                  || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

```
RESULT   8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;     APPLICANT:  Hoffman, Stephen L.
;     APPLICANT:  Charoenvit, Yupin
;     APPLICANT:  Hedstrom, Richard
;     APPLICANT:  Khusmith, Srisin
;     APPLICANT:  Rogers IV, William O.
;     TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;     TITLE OF INVENTION:  immunogen and gene encoding
;     NUMBER OF SEQUENCES:  2
;     CORRESPONDENCE ADDRESS:
;       ADDRESSEE:  A. David Spevack
;       STREET:  NMRDC Building 1 T-12 National Naval
;       STREET:  Medical Center
;       CITY:  Bethesda
;       STATE:  MD
;       COUNTRY:  USA
;       ZIP:  20814-5044
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.24
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  PCT/US92/00018
;       FILING DATE:  19920103
;       CLASSIFICATION:  424
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Spevack, Avram D.
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (301) 295-6759
;       TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  826 amino acids
;       TYPE:  AMINO ACID
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  protein
PCT-US92-00018-2
```

```
  Query Match             6.7%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 3.9e-23;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    324 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 376
```

```
          ||  ::  |  |  |  |      |: |:||  |    :|     |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNN 332

Qy    377 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 432
          |      : |  |  |  |  |  |  |  |  |  |  |  |  |  |  | :|:     |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    433 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 490
            ||   |  ||  || |  |     |: | :| |       |: | :| |    : |:
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    491 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 548
            :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    549 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 607
          |:         |  |  |  |: | |  |: |||                    |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    608 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 666
          : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    667 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 726
                    |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    727 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 786
          ||||                |  : ||:|    : |   : |     :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    787 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 840
          |  |             |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    841 PTPKALENSPK------------------EPGVPTTK 859
          |     |  ||                  || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

```
RESULT   9
5202236-13
;Patent No. 5202236
;     APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;     TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;     NUMBER OF SEQUENCES: 39
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER: US/07/528,762
;       FILING DATE: 25-MAY-1990
;       APPLICATION NUMBER: 82,456
;       FILING DATE: 07-AUG-1987
;       APPLICATION NUMBER: 933,945
;       FILING DATE: 24-NOV-1986
;       APPLICATION NUMBER: 650,128
;       FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;       LENGTH: 652
5202236-13
```

```
  Query Match             6.5%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 1.7e-22;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy    207 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 266
          ||:|      |:|| ||                | ::| |              | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    267 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 322
                  || |  |||:          |    ||:  | | |     | :  ||
Db     41 -----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79
```

```
Qy    323 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 379
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    380 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 434
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196

Qy    435 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 485
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    486 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 532
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308

Qy    533 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 588
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    589 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 643
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    644 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 696
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    697 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 754
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    755 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 807
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    808 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 867
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    868 MTTTAKDKTTERDLRTTPETTTAAP 892
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645


RESULT 10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
;  APPLICANT:  VAN HEERDE, GEORGE V.
;  APPLICANT:  VAN RIJN, ALEXIS C.
;  APPLICANT:  BOUWSTRA, JAN B.
;  APPLICANT:  DE WOLF, FREDERIK A.
;  APPLICANT:  MOOBROEK, ANDREAS
;  APPLICANT:  WERTEN, MARC W.T.
;  APPLICANT:  WIND, RICHELE D.
;  APPLICANT:  VAN DEN BOSCH, TANJA J.
;  TITLE OF INVENTION:  SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
;  TITLE OF INVENTION:  SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
;  TITLE OF INVENTION:  PREPARATION THEREOF
;  FILE REFERENCE:  2728-2
;  CURRENT APPLICATION NUMBER: US/09/219,849
;  CURRENT FILING DATE:  1998-12-23
;  NUMBER OF SEQ ID NOS:  50
;  SOFTWARE:  PatentIn Ver. 2.1
; SEQ ID NO 5
;  LENGTH: 960
;  TYPE: PRT
;  ORGANISM: Artificial Sequence
;  FEATURE:
;  OTHER INFORMATION: Description of Artificial Sequence: Illustrative
;  OTHER INFORMATION: amino acid sequence
US-09-219-849-5


  Query Match             6.2%;  Score 455.5;  DB 4;  Length 960;
  Best Local Similarity   23.0%;  Pred. No. 5.5e-21;
  Matches  195;  Conservative   64;  Mismatches  338;  Indels  251;  Gaps   46;

Qy    110 PTSPPSSKKAPPPSGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKKKP 169
Db     23 PAGPPGSRD-PGPPGAPGPAGPPGSRDPGPPG-------------------------AP 55

Qy    170 TPK-PPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTS 228
Db     56 GPAGPPGSRDPG-------------------------------PPGAPGPAGPPGS--- 80

Qy    229 KETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKP 288
Db     81 ---------------------------------------------RDPGPP----GAP 89

Qy    289 TPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPT 348
Db     90 GPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGP------PG 138

Qy    349 TPKEPAPTTTKEPAPTTPKEPA-PTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP-- 405
Db    139 APGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPG 198

Qy    406 -KEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPA---------------------- 440
Db    199 SRDPGP--PGAPGPAGP----PGSRDPGPPGAHGPAGPKGAHGPAGPKGAH 252

Qy    441 -PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKSAPTTTKEP 494
Db    253 GPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----AGPPGSRDP 306

Qy    495 APTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAPTTP---KE 548
Db    307 GP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRD 356

Qy    549 PAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT- 600
Db    357 PGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPGSRDPGPPG 412

Qy    601 TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPTTPKEPA-- 654
Db    413 APGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGP 470

Qy    655 -----PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELAPTTTKEPT 699
Db    471 KGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 530

Qy    700 STTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEPAPTTPKKP 754
Db    531 PPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAP 586

Qy    755 AP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTP-- 804
Db    587 GPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPG 644

Qy    805 --ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----SPKEPGVPT 857
Db    645 SRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPG 703

Qy    858 TKTPAATK 865
Db    704 AHGPAGPK 711


RESULT 11
US-09-041-886-23
; Sequence 23, Application US/09041886
; Patent No. 6235872
```

```
;  GENERAL INFORMATION:
;    APPLICANT:  Bredesen, Dale E.
;    APPLICANT:  Rabizadeh, Sharroz
;    TITLE OF INVENTION:  Proapoptotic Peptides, Dependence
;    TITLE OF INVENTION:  Polypeptides and Methods of Use
;    NUMBER OF SEQUENCES:  72
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Campbell & Flores LLP
;      STREET:  4370 La Jolla Village Drive, Suite 700
;      CITY:  San Diego
;      STATE:  California
;      COUNTRY:  United States
;      ZIP:  92122
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.25
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/041,886
;      FILING DATE:
;      CLASSIFICATION:
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Campbell, Cathryn A.
;      REGISTRATION NUMBER:  31,815
;      REFERENCE/DOCKET NUMBER:  P-LJ 2626
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (619) 535-9001
;      TELEFAX:  (619) 535-8949
;  INFORMATION FOR SEQ ID NO:  23:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1185 amino acids
;      TYPE:  amino acid
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-09-041-886-23


  Query Match             5.8%;  Score 428;  DB 4;  Length 1185;
  Best Local Similarity   22.1%;  Pred. No. 3.7e-19;
  Matches  237;  Conservative  114;  Mismatches  429;  Indels  292;  Gaps   50;

Qy     117 KKAPPP-------SGASQTIKSTTKRSPKPPNKKKTKKVIESEEITEVKDNKKNRTKKKP 169
           |:|| |         ||    ||:    |    ::| |     || :  | ||: |:::
Db      18 KEAPGPREELRSRGRASPGGVSTSSSDGKAEKSRQTAKKARVEEASTPKVNKQGRSEE-- 75

Qy     170 TPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSK 229
           : |: |                    : : :|| |  :   |||  | : |:
Db      76 ------ISESES-------------------EETNAPKKTKTEQELPRPQSPSDLDSLD 109

Qy     230 ETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPT 289
           ||  | : : : ::    |: ||        |     |:|  |
Db     110 GRSL--NDDGSSDPRDIDQDNRSTS-------PSIYSPGSVENDS-------------- 145

Qy     290 PKAETTTKGPALTTPKEPTPTTP--KEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 347
           :   ::|||   |  | | |  |  :|  :||::|  :          || |: | :
Db     146 DSSSGLSQGPA--RPYHPPPLFPPSPQPPDSTPRQPEASF---------EPHPSVTPTGY 194

Qy     348 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPT--TTKSAPTTPK-------------- 390
           | || ||:   ||  |  | |     |  |       |  ||
Db     195 HAPMEP-PTSRMFQAP--PGAPPPHPQLYPGGTGGVLSGPPMGPKGGGAASSVGGPNGGK 251

Qy     391 -EPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAP-----TTPKEPAPTA- 443
           | ||||   : :      || |   ||||          | |      ||  | | |
Db     252 QHPPPTTPISVSSSGASGAPPTKP----PTTPVGGGNLPSAPPPANFPHVTPNLPPPPAL 307

Qy     444 -PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE------PAPTTTKSAPTTTKEP-- 494
           |   |  :|    |   :| |     |||    :      | |   :  ||    |
Db     308 RPLNNASASP----PGLGAQPLPGHL--PSPYAMGQGMGGLPPGP---EKGPTLAPSPHS 358

Qy     495 APTTTKSAPTTP-KEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE-PAPT 552
           |  : |||  | : |  :::   |  :    : ::   | | :  :  |: |   | ||
Db     359 LPPASSSAPAPPMRFPYSSSSSSSAAASSSSSSSSSSASPFPAS--QALPSYPHSFPPPT 416

Qy     553 TTKKPAPTAPKEPAPTTPKETAPTTPKKLT-----PTTPE--KLAPTTPEKPAPTTPEEL 605
           :      :   :|    || | |: | :       |  |   :|   :   | |  |
Db     417 SL-----SVSNQP----PKYTQPSLPSQAVWSQGPPPPPPYGRLLANSNAHPGPFPPSTG 467

Qy     606 APTTPEEPTPT----------------------TPEEPAP-----------TTPKAA 629
           | :|   |  |                      |  |                 | |
Db     468 AQSTAHPPVSTHHHHHQQQQQQQQQQQQQQQHHGNSGPPPPGAFPHPLEGGSSHHAHPYAM 527

Qy     630 APN----TPKEPAPTTPKEP------APTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAP 679
           :|:     |  | |     |       :   |  |  ::   :: :| :|: | :   |:|
Db     528 SPSLGSLRPYPPGPAHLPPPHSQVSYSQAGPNGPPVSSSSNSSSSTSQGSYPCS--HPSP 585

Qy     680 TTPKKPAPKELAP---TTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPK 736
           :   : ||    |    ||   | :|      ::| |    :|  | |   : |:|   |
Db     586 SQGPQGAPYPFPPVPTVTTSSATLSTVIATVASSPAGYKTASPPGPPPYGKRAPSPGAYK 645

Qy     737 GTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTP 796
             |  | :|    |: :  |   :|    ||    | | |  :||    | |  |
Db     646 TATPPGYKPGSP-----PSFRTGTPPGYRG----TSPPAGPGTFKPGSPTV--GPGPLPP 694

Qy     797 KKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE--STPELSAEPTPKALENS--PKE 852
             |:   |  |||  :  | |  :   | | : | |   |||    | | |   |  ||
Db     695 AGPS-GLPSLPPPPAAPASGPPLS--ATQIKQEPAEEYETPE---SPVPPARSPSPPPKV 748

Qy     853 PGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKT-TESKIT 911
             ||:  : :|   :      :    ||   |       |: |: |   ||   |::
Db     749 VDVPSHASQSARFNKHLDRGFNSCARSDLYFVP---LEGSKLAKKRADLVEKVRREAEQR 805

Qy     912 ATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNS 971
           |                                                ||  : ::|
Db     806 AR----------------------------------------------EEKERERERERERK 820

Qy     972 KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQ 1031
           :    | ::  :: |     : |     |  |   | | :     |   | | :|
Db     821 ERERERERELERSVKLAQEGRAP-----VECPSLGPVPHR-----PPFEPGSAVA----- 865

Qy    1032 TTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLP 1083
           |  |   |::  :     || |           |||| |       |:|
Db     866 -TVPPYLGPDTPALRT--LSEYA----------RPHVMSPGNRNHPFYVP 902


RESULT  12
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
```

```
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5

  Query Match             5.8%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 1e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    231 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 289
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    290 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 332
          |    :||:|    |    ||               | |     ||||     |   ||||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    333 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 373
              :||| :|        |||     |            ||     |  | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    374 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 420
          :  | | :|         |||     |  | |       ||     |  | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    421 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 467
          :  | | :  |          ||     |  |||       ||     |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    468 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 516
          :  |  |   ||      |||      |  |                |||     |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    517 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 568
          ||      |  | |        ||      |  | |       ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    569 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 624
          :  :  |    | |     |:    ||      |  | |     |:  :  |     | |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    625 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 679
          | |     :||:     |  | |      |:  :  | |  |||      | ||      |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    680 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 727
          | |           :| |   ||      ||    :| |  | |      |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    728 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 773
          |  | |       :||     |  | |                :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    774 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 817
          :| |  | |       :||     |  | |                :| | :  | |     | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    818 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 863
              | :       :  :||      |  |    : ::|     | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    864 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 919
           :|   :||      |:  | |  | :|  ||     |:    :                |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    920 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 979
                |    |  |  ||      : :   |  ||     |::   | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    980 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1039
           : ||     |  |   |      | |    :||  |:  |                | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1040 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1098
           | :  |   |       |:     |          ||   |  |  :     |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1099 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1153
           |:  : :|  | |    :|   |          |   | |         |: | |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003

RESULT  13
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
```

```
;       ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;       STREET:  P.O. Box 1404
;       CITY:  Alexandria
;       STATE:  Virginia
;       COUNTRY:  United States
;       ZIP:  22313-1404
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.0, Version #1.30
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  US/09/083,116
;       FILING DATE:
;       CLASSIFICATION:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  08/479,537
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5
```

```
  Query Match             5.8%;  Score 424.5;  DB 4;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 1e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    231 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 289
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    290 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 332
          |   :||:|   |    ||                | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    333 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 373
          :||| :|         |||      |              ||     | | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    374 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 420
            : | | :|         |||      | | |        ||      | | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    421 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 467
            : | | :  |         ||      |  |||        ||      | | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    468 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 516
            :  |  |   ||       |||      | |          |||     |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    517 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 568
          ||      | | |        ||      | | |        ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    569 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 624
            :  :  |    | |      |:    ||      | | |      |:  :  |     |  |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    625 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 679
            | |     :||:    |  | |      |:  :  |  |  |||      | ||     |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    680 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 727
          | |          :| |   ||       ||     :| |  | |      |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    728 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 773
            |  | |       :||      |  | |          :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    774 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 817
            :| |  | |       :||      |  | |          :| | :  | |       | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    818 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 863
              | :       :  :||      |  |     : ::|      | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    864 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 919
            :|   :||       |:  | |  | :|  ||      |:    :              |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    920 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 979
                |       |   | ||      :  :    |  ||     |::    | | `
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    980 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1039
            : ||      |  |   |       | |    :||  |:  |                |  |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1040 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1098
            | :  |   |     |:      |           ||   |  |   :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957
```

```
Qy   1099 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1153
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  14
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2


  Query Match             5.8%;  Score 424.5;  DB 2;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 1.1e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    231 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 289
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    290 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 332
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    333 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 373
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    374 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 420
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    421 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 467
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    468 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 516
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    517 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 568
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    569 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 624
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    625 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 679
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    680 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 727
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    728 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 773
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    774 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 817
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    818 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 863
```

```
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    864 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 919
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    920 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 979
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    980 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1039
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1040 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1098
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1099 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1153
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

```
RESULT  15
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repea
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CC
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2
```

```
  Query Match             5.8%;  Score 424.5;  DB 4;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 1.1e-18;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    231 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 289
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    290 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 332
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    333 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 373
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    374 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 420
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    421 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 467
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    468 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 516
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    517 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 568
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    569 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 624
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471
```

```
Qy    625 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 679
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    680 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 727
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    728 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 773
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    774 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 817
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    818 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 863
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    864 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 919
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    920 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 979
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    980 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 1039
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy   1040 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1098
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1099 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1153
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

Search completed: April 26, 2002, 16:26:24
Job time: 524 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:32:49 ; Search time 114.61 Seconds
(without alignments)
871.345 Million cell updates/sec

Title: AA8
Perfect score: 7064
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1311

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1078.5	15.3	3020	2	A43932	mucin 2 precursor,
2	950	13.4	1664	2	T18262	S-layer protein -
3	855	12.1	1489	2	T31108	cyst germination s
4	802	11.4	1274	2	T16251	hypothetical prote
5	798.5	11.3	1367	1	S48478	glucan 1,4-alpha-g
6	784	11.1	1188	2	S49915	extensin-like prot
7	776.5	11.0	2187	2	T30826	nascent polypeptid
8	672	9.5	1344	1	A35175	mucin 1 precursor,
9	671	9.5	1151	2	T18535	high molecular mas
10	662	9.4	1229	2	T25697	hypothetical prote
11	661	9.4	3507	2	T34513	hypothetical prote
12	646.5	9.2	3570	2	T45025	mucin MUC5B, trach
13	635.5	9.0	7962	2	I38346	elastic titin - hu
14	633	9.0	489	2	T11622	extensin class 1 p
15	632	8.9	761	2	C84672	hypothetical prote
16	630.5	8.9	990	2	I51618	nucleolar phosphop
17	626.5	8.9	971	2	T19431	hypothetical prote
18	625	8.8	6642	2	T29757	protein UNC-89 - C
19	622.5	8.8	839	2	F75518	hypothetical prote
20	608	8.6	801	2	T29018	hypothetical prote
21	607.5	8.6	924	2	S27923	gene LF3 protein -
22	605	8.6	379	2	S50125	larval glue protei
23	559.5	7.9	350	2	S22456	hydroxyproline-ric
24	556.5	7.9	856	2	T16543	hypothetical prote
25	556.5	7.9	875	2	S23760	polyphenolic adhes
26	555.5	7.9	1630	2	A53577	ascites sialoglyco
27	554	7.8	2232	2	T34434	hypothetical prote
28	551	7.8	620	2	S06733	hydroxyproline-ric
29	545	7.7	873	2	A47283	calphotin - fruit
30	542	7.7	369	2	S20500	hydroxyproline-ric
31	538.5	7.6	416	2	JU0465	extensin precursor
32	533	7.5	1087	1	QFMSH	neurofilament trip
33	532	7.5	756	2	T27642	hypothetical prote
34	530.5	7.5	1162	2	JH0557	exo-alpha-sialidas
35	530	7.5	865	2	A47282	calcium-binding pr
36	522	7.4	328	2	JQ0985	hydroxyproline-ric
37	522	7.4	1459	2	T32271	hypothetical prote
38	518	7.3	813	2	S70795	vsaA protein precu
39	518	7.3	866	2	T45462	membrane glycoprot
40	518	7.3	1072	1	A37221	neurofilament trip
41	512.5	7.3	867	2	T45463	membrane glycoprot
42	509.5	7.2	662	2	A45155	mucin FIM-C.1 - Af
43	508	7.2	1832	2	T31113	mucin-like glycopr
44	505	7.1	700	2	A54641	interspersed repea
45	504.5	7.1	606	2	A43427	neurofilament trip

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 15.3%; Score 1078.5; DB 2; Length 3020;
Best Local Similarity 28.1%; Pred. No. 9.1e-39;
Matches 378; Conservative 88; Mismatches 478; Indels 399; Gaps 52;

```
Qy      43 TCNCDY--NCQHYMECCPDFKRVCTAELSC------------------KGRC---FE--- 76
           :|:||   :|: :      : : || | :|                    |   :|
Db    1079 SCSCDTGGDCECFCSAVASYAQECTKEGACVFWRTPDLCPIFCDYYNPPHECEWHYEPCG 1138

Qy      77 --SFERGRECD-------------CDAQCKKYDKCCPDYESFCAEVKDNKKNRTKKK--- 118
             |||  |  :             |  :| | |:  | ||      :| ||  |  |
Db    1139 NRSFETCRTINGIHSNISVSYLEGCYPRCPK-DR--PIYE------EDLKKCVTADKCGC 1189

Qy     119 ---PTPKPP-------------------VV--DEAGSGLD---NGDF---KVTTPDTST 147
              |  ||                    ||   | |  |:   :| |   ::  |: :
Db    1190 YVEDTHYPPGASVPTEETCKSCVCTNSSQVVCRPEEGKILNQTQDGAFCYWEICGPNGTV 1249

Qy     148 TQH---NKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKET--TTTNK- 201
            :|     ::| |   |       |: |       || |    ||  |  |  :|| |
Db    1250 EKHFNICSITTRPSTLTTFTTITLPTTP--------TSFTTTTTTTTPTSSTVLSTTPKL 1301

Qy     202 --------------QTSTDGKEKTTS--AKETQSIEKTSAKD------------------ 227
                         | ||  :         : ||  | ||
Db    1302 CCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSV 1361

Qy     228 ---------------------------------LAPTSKVLAKPTPKAETTTKGPALTT 253
                                             :   |     |:|   |||  |  ||
Db    1362 GFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTT 1421

Qy     254 PKEPTPTTPKEPASTTPKEPTPTTIKSAP-TTPKEPAPTTTKSAPTTPKEPAPTTTKEPA 312
           |  || ||   | :|||  |  ||    | |||  |  |||   |||   | ||||  |
Db    1422 PSPPTTTTTTPPPTTTPSPPITTTTTPLPTTTPSPPISTTTTPPPTTTPSP-PTTTPSPP 1480

Qy     313 PTTPKEPAPTTTKEPAPTTTKSAP-TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTP 371
             |||  |  |||  | |||| | | |||  |  :|   |  |||  |  ||   |  |||
Db    1481 TTTPSPPT-TTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTTTTPPPTTTP 1539

Qy     372 KEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAP 431
             |  |   | ||:     ||    | |||   | |||   |  |||  |   |   | |
Db    1540 SPPTTTPITP-PTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPTITTTTPPP 1598

Qy     432 TTTKSAP-TTTKEPAPTTTKSAP-TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPK 489
           ||| | | |||  | |||| | | |||  |  :||  |  |||  |  ||   |  |||
Db    1599 TTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPS 1658
```

```
Qy     490 EPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAP 549
           |  |||  |  |||  |  | |  |  |||  : |||    | |||     |||  |
Db    1659 PPTTTTPSPPITTTTTPPPTTTPSSPITTTP--SPPTT----TMTTPS--PTTTPSSPIT 1710

Qy     550 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTP- 608
           ||    : |||  | |||   |:|||  :    |     |||   |  |||  |: | |
Db    1711 TTTTPSSTTTP-SPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPT 1769

Qy     609 ----KETAPTTP-----------KG---------------------------------- 618
                 | ||||              |
Db    1770 FSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYP 1829

Qy     619 ------------------------------------------------TAPTTLKEPA 628
                                                           | |||:
Db    1830 DVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTM---T 1886

Qy     629 PTTPKKPAPKELAPTT--TKEPTSTTSDKP----APT-TPKGTAPTTPKEPAPTTPKEPA 681
           || : | |  :  ||  |  || |::  |    |||  ||  |  |    | ||  : |
Db    1887 TTTTENPTPTPITTTTTVTPTPTPTSTQSPNGLQAPTPTPISTTTTVTPTPTPTGTQTPT 1946

Qy     682 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKPAPTTPKETAPTTPKE 740
            |||  |  |    | ||  : |    :  |||  || | || |    ||  |  |
Db    1947 -TTPITTTTTVTPTPTPTGTQTPTTVLITTTTTMTPTPTPTSTKSTTVTPITTTTTVTAT 2005

Qy     741 PAPTTPKKP-------APTTPETPPPTTSEVSTPTTTK-----EPTTIHKSPDESTPELS 788
           | ||  : |          |   || |||      || |       | ||  :  |: |||: |
Db    2006 PTPTGTQTPTMIPISTTTTVTPTPTPTTGSTGPPTHTSTAPIAELTTSNPPPESSTPQTS 2065

Qy     789 AEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKET 848
             :    |::     :|      :| | | ||   |:     : | :|| :|  |
Db    2066 RSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPTTTSGGHTLSPPPSTTTSPPGTPTR 2125

Qy     849 ATTTEKTTESKITATTTQVTSTTTQDTTPFKITT---LKTTTLAPKVTTTKKTITTTEIM 905
           |||   : |  | :| | |:|:   ||  ::|   ::|| | |  ::        :
Db    2126 GTTT--GSSSAPTPSTVQTTTTSAWTPTPTPLSTPSIIRTTGLRPYPSSVLICCVLNDTY 2183

Qy     906 NKPEE----------------------------TAKPKDRATNSKATTPKPQKPTKAPK 936
             | |                          : |    | || :|| | ||:  |
Db    2184 YAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTPSPTPTPSK-STPTPSKPSSTPS 2242

Qy     937 KPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVE 996
           |||   ||   |   |:             : | |  : :  :  | : | |     ||
Db    2243 KPTPGTKPPECPDFDPPR------------QENETWWLCDCFM-ATCKYNNTVEIVKVE 2288

Qy     997 VNP------------KSEDAGG-----------AEGETPHMLL----------RPHV 1020
             |             : ||  |          |: || :           :|
Db    2289 CEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGD-PHYVTFDGLYYSYQGNCTYV 2347

Qy     1021 FMPEVTPDMDYLPRVPNQGIIIN 1043
             : |::| :|      | |: |:
Db    2348 LVEEISPSVD------NFGVYID 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

Query Match 13.4%; Score 950; DB 2; Length 1664;
Best Local Similarity 31.0%; Pred. No. 1.6e-33;

Matches 312; Conservative 104; Mismatches 361; Indels 230; Gaps 55;

```
Qy    234 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 292
Db    758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy    293 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 348
Db    805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy    349 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 407
Db    865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy    408 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 460
Db    920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy    461 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 515
Db    980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy    516 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 568
Db   1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy    569 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 625
Db   1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy    626 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 684
Db   1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy    685 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 740
Db   1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy    741 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 799
Db   1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy    800 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 857
Db   1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy    858 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 917
Db   1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT----------------PTSG 1396

Qy    918 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 962
Db   1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy    963 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 1014
Db   1454 YLRGYPDGSFRPERNITRAEAAVIF----------AKLL----GADESYGAQSASPYSD 1498

Qy   1015 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 1040
Db   1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy   1041 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1087
Db   1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT----------FKPQN 1600

Qy   1088 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1132
Db   1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

RESULT 3
T31108

cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90

Query Match 12.1%; Score 855; DB 2; Length 1489;
Best Local Similarity 31.8%; Pred. No. 1.6e-29;
Matches 362; Conservative 52; Mismatches 496; Indels 228; Gaps 51;

```
Qy     78 FERGRECDCDAQCKKYDKCCPDYESFCAEVKDN---KKNRTKKKPTPKPPVVDEAGSGLD 134
Db    301 YEPSDETEAPTEGTTY---VPREETTAAPSEDTTYAPREVTPYAPTEKPYDVEETTYVTE 357

Qy    135 NGDFKVTTPDTSTT------QHNKVSTSPKITTAKPINPRPSLP--------------P 173
Db    358 ESTYAPTKSETNAPTERMHYAHIEKPCDTEVTMYAPTEETTYAPTEETTYAPTEETTYAP 417

Qy    174 NSDT----SKETSLTVNKETTVETKETTT---TNKQTSTDGKEKTTSAKETQSIEKTSAK 226
Db    418 TEETPYEPTEETTYTPTEETTYAPTEETTYAPTEKTTYAPTEETTYAPTEETPYEPTEET 477

Qy    227 DLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPK 286
Db    478 TYAPTKETTYAPT---EETTYASTEETTYAPTEETTYAPAEETPYEPTEET-TYAPTEET 533

Qy    287 EPAPT-TTKSAPTTPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPT------ 337
Db    534 TYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETT-YAPTEETMYA 592

Qy    338 ----TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEP--------------AP--- 376
Db    593 PIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYASTEETTYAPTEE 652

Qy    377 TTKEPAPTTPKEP------APTAPKKPAPTTPKEPAPTTPKEPAPT--TTKEPSPTTPKE 428
Db    653 TTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYE 712

Qy    429 P------APT-TTKSAPT----------TTKEP------APT-TTKSAPT--TPKEPSPT 462
Db    713 PTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTYAPTEATTYAPTEETPYAPTEE 772

Qy    463 TTKEP------APTTPKEPAPTTPKKPAPTTPKEPAPT--TPKEPAPTTTKKPAPTAPKE 514
Db    773 TTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPTEETTYAPTEETPYE 832

Qy    515 PAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT--TPEEPTPTTPEEP- 571
Db    833 PTEET--TYTPTEETTYAPTEETTYAPTEKTTYAPTEETTYAPTEETPYEPTEETTYAPT 890

Qy    572 -----APT----------TPKAAAPNTPKEPAPTTPKEP------APTTPKEPAPTTPKE 610
Db    891 KETTYAPTEETTYASTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETT 950

Qy    611 TAPTTPKGTAPT--TLKEPAPTTPKKPAPK-ELAPT--TTKEPTSTTSDKPAPTTPKGTA 665
Db    951 YAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYAPT 1010

Qy    666 PTTPKEPAPTTPKEP------APTTPKGTAPT----------------TLKEPAPTTP 701
Db   1011 EETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYASTEETTYAPTEETTYAPAEETP 1070

Qy    702 KKP------APKE---LAPT--TTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPA 750
```

```
Db   1071 YEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPT 1130

Qy    751 PTTPETPPPTTSEVSTPTTTKEPT--TIHKSPDESTPELSAEPTPKA-LENSPKEPG--- 804

Db   1131 EETTYAPTEETMYAPIEETTYGPTEETTY-APTEATTYAPTEETPYAPTEETTYEPTGET 1189

Qy    805 --VPTTKTPAATKPEMTTTAKDKTTERDLRTTP----ETTTAAPKMTKETATTTEKTTES 858

Db   1190 TYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPTEETTYAP-----TEETTYEPTEE 1244

Qy    859 KITATTTQVTSTTTQDTT----------PFKITT---LKTTTLAPKVTTTKKTITTTEIM 905

Db   1245 TTYAPTEETTYAPTEETTYAPTEETMYAPIDETTYGPTEETTYAPTEATTYAPTEETPYA 1304

Qy    906 NKPEETAKPKDRAT--NSKATTPKP-QKPTKAPKKPT-----------STKKP-KTMPR 949

Db   1305 PTEETTYEPTGETTYAPTEETTYAPTEETTYAPMEETPYEPAEESTSTVSTEKPCNTEEF 1364

Qy    950 VRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP--NQTPNSKLVEVNPKSEDAG 1005

Db   1365 TDEPTDEPTDE--PSDEPTDEPTDEPTDLPTDEPSTPCDNQGINGIGVENKVRYNNAG 1420
```

RESULT 4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN0
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

Query Match 11.4%; Score 802; DB 2; Length 1274;
Best Local Similarity 28.0%; Pred. No. 2.5e-27;
Matches 290; Conservative 102; Mismatches 414; Indels 230; Gaps 56;

```
Qy    112 KNRTKK-KP-----TP-----KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTS 156

Db    274 KNPTKKWKPPWEDETPVEEVKEPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPS 333

Qy    157 PKITTAKPI-NPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 215

Db    334 SPVVPPTPVKNPVKKYKPPWEVDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPS 393

Qy    216 ETQSIEKTSAKDLAPTSKVLAKPTPKAE----TTTKGPA----------------LTTP 254

Db    394 PKKAAPAVKPRDSSPKKATPLQADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQP 453

Qy    255 KEPTPTTP----KEPASTTPKEP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAP 305

Db    454 EAPAKKTPVLKRKEPAAKDTAKPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA- 512

Qy    306 TTTKEPA------------------PTTPKEPAPTTTKEPAPT-------TTKSAPTTP 339

Db    513 TPVKNPVKKWRPPWEDDETPADDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKP 572

Qy    340 KEP-----APTTPKKPAPTTPKE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EP 389

Db    573 RDPSPKKVAPTAPEKKTPVLAKKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEV 632

Qy    390 APTAPKKPAP------TTPKEPAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPT 439

Db    633 APAAVKKPEPISKPKDTAPKKAEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPE 691

Qy    440 TTKEPAPTTTKSAPTTP-KEPSPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPT 494

Db    692 PEKK-TPVLAKKAPTKPDSEAAADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPK 750

Qy    495 TPKEPAPTTTKKPAPTA-PKEPAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPA 548

Db    751 T--EVPPAVVKKPEPVAKSRDPSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDA 805

Qy    549 PTTPEELAPTTPEEP-----TPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKE 602

Db    806 PAEPVNVPEPEKKTPVLAKKTPVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKE 865

Qy    603 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTP 661

Db    866 PSPKKAEPNSPVVP----PTPVKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLA 918

Qy    662 KGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPT-------TPKKPAP-----KEL 709

Db    919 K-KEPEKPKD-APKVAAKPRDPSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEA 974

Qy    710 APTTTKGPTS------TTSDKP-----APTTPKET------APTTPKEPAPTTPKKPAPT 752

Db    975 PPTPVKNPVKKWKPPWEDDDEPAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPV 1031

Qy    753 TPETPPPTTSEVSTPTTTKEPTTIHKSP----DESTPELSA-EP---TPKALENSPKEPG 804

Db   1032 AAK-PDPKIPEV-PPTPVKNPVKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPA 1089

Qy    805 V-------------PTTKTPAAT------KPEMTTTAKDKTTERDLRTTPETT--TAAPK 843

Db   1090 TKPDSEAAADPVSGPTSKDPKLSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPK 1149

Qy    844 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 903

Db   1150 KWKPVWDDDPDEPEADFTVPAPSKKPDTEDPADPLG---------GPKTKDPK------- 1193

Qy    904 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK--------- 954

Db   1194 -LNKKAPAEKPTEK--------PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPE 1243

Qy    955 ---TTPTPRKMTSTMP 967

Db   1244 ADFTMPAPKKPDTEDP 1259
```

RESULT 5
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877

```
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacchar
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseudohy
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosidase
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>

  Query Match           11.3%;  Score 798.5;  DB 1;  Length 1367;
  Best Local Similarity  27.9%;  Pred. No. 3.7e-27;
  Matches  313;  Conservative  108;  Mismatches  523;  Indels  177;  Gaps   50;

Qy      85 DCDAQCKKYDK---CCPDYESFCAEVK-DNKKNRTKKKPTPKPPVVDE--AGSGLDNGDF 138
Db     174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSESSTTTSSTSESSTT 233

Qy     139 KVTTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETT 190
Db     234 TSSTSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTT 293

Qy     191 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKG 248
Db     294 PCTKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSA 342

Qy     249 PALTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPT 298
Db     343 PVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSST 402

Qy     299 TPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP 358
Db     403 TESSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTT 451

Qy     359 APTTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPK 411
Db     452 ESSSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTES 507

Qy     412 EPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEP 467
Db     508 SSAPVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 566

Qy     468 AP----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKE 522
Db     567 TPVTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESS 622

Qy     523 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTP 576
Db     623 SAPVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 677

Qy     577 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKP 635
Db     678 SSSAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSST 729

Qy     636 APKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPT 691
Db     730 TESSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 788

Qy     692 TLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTP 746
Db     789 SSSAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTE 848

Qy     747 KKPAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPK 801
Db     849 SSSAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSK 904

Qy     802 EPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKIT 861
Db     905 YPGSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVT 964

Qy     862 AT--TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPK 915
Db     965 TTVPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--- 1021

Qy     916 DRATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPT 958
Db    1022 -----SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPT 1076

Qy     959 PRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRP 1018
Db    1077 P--SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGCSPKTVTTTVPC 1120

Qy    1019 HVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFW 1078
Db    1121 STGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVT 1173

Qy    1079 MLSPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1115
Db    1174 TLAPSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207


RESULT   6
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

  Query Match           11.1%;  Score 784;  DB 2;  Length 1188;
  Best Local Similarity  29.9%;  Pred. No. 1.3e-26;
  Matches  230;  Conservative   59;  Mismatches  359;  Indels  122;  Gaps   33;

Qy     119 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 174
Db     460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy     175 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 234
Db     511 APVG-----------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy     235 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 289
Db     542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy     290 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 346
```

```
Db    600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy    347 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 397

Db    660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy    398 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 457

Db    719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy    458 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 517

Db    769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy    518 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 564

Db    828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy    565 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 623

Db    886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy    624 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 669

Db    944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000

Qy    670 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 728

Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    729 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 787

Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    788 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 837

Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158
```

```
RESULT   7
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a muscle
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor


  Query Match             11.0%;  Score 776.5;  DB 2;  Length 2187;
  Best Local Similarity   27.6%;  Pred. No. 5e-26;
  Matches  290;  Conservative  119;  Mismatches  416;  Indels  227;  Gaps   48;

Qy     98 PDYESFCAEVKDNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNK----- 152

Db    809 PKVDPIMSDVTPTSPKKTSATAVPK-------------------DTSATLSLKSVPAV 847

Qy    153 VSTSPKITTAKPINPRPSLPPNSDTSKETSL--TVNKET------TVETKETTTTNKQTS 204
```

```
Db    848 TSLSPPKAPVAPSNEATIVPTEIPTSLKNALAAATPKETLATSIPKVTSPSPQKTPKSVS 907

Qy    205 TDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE 264

Db    908 LKGAPAMTSKKAT---EIAASKDVSPSQ--FPKEVPLLQHV---PPTSPPKSPVSDTLSG 959

Qy    265 PASTTPKEPTPTTIKSAPTTPKE-PAPTTTKSAPTTPK---------EPAPTTTKEPAPT 314

Db    960 ALTSPPPKGPPATLAETPTYPKKSPKPAASKKTPATPSPEGVTAVPLEIPPCSKKAPKTA 1019

Qy    315 TPKEPAPTTTKEPAPTT--TKSAPT--------------------TPKEPAPTTPKK-P 350

Db   1020 APKESSATSSSKRAPKTAVSKEIPSKGVTAVPLEISLPLKETSKSATPGEKSASSPKRSP 1079

Qy    351 APTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTT--PKEPAP- 407

Db   1080 KTAGPKE---TPPGGVTAVPPEISLPPKETPQNATPNESLAASSQKRSPKTSVPKETPPG 1136

Qy    408 ---TTPKEPAPTTTKEPSPTTPK-----EPAPTTTKSAPTTTKEPAPTTT-KSAPTTPK- 457

Db   1137 GVTAMPLEIPSAPQKAPKTAVPKQIPTPEDAVTILAGSPLSPKKASKTAAPKEAPATPSV 1196

Qy    458 -------EPSPTTTKEPAPTTPKEPAPTTPKKPAPTT--PKE------------PAPTTP 496

Db   1197 GVIAVSGEISPSPKKTSKTAAPKENSATLPPKRSPKTAAPKETPATSSEGVTAVPSEISP 1256

Qy    497 KEPAP-------TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK-PA 548

Db   1257 SPPTPASKGVPVTLTPKGAPNALAE-SPASPKKVPKTAAPEETSTTP------SPQKIPK 1309

Qy    549 PTTPEELAPTTPEEPTPTT--------PEEPAPTTPKAAAPNTPKEPAPTTPKE-PAPTT 599

Db   1310 VAGPKEASATPPSKKTPKTAVPKETSAPSEGVTAVPLEIPPSPRKAPKTAAPKETPAPS- 1368

Qy    600 PKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP-------------------KKPAPKE 639

Db   1369 -PEGATTAPVQIPPSPRKGSKKAGSKE-TPTTPSPEGVTAAPLEIPISSKKTSKMASPKE 1426

Qy    640 LAPTTTKEPTS------TTSDKPAPTTPKGTAPTTPKEPAPTTPKE-PAPTTPKGTAPTT 692

Db   1427 TLVTPSSKKLSQTVGPKETSLEGATAVPLEIPPSHKKAPKTVDPKQVPLTPSPK-DAPTT 1485

Qy    693 LKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPT 752

Db   1486 LAE-SPSSPKK-APKTAAPPSER-VTTVPPEKPA-TPQKASGTTASKVPVPAETQEVAVS 1541

Qy    753 TPETP-----PPTTSEVSTPTTTK------EPTTIHKSPDESTPELSAEPTPKALENSPK 801

Db   1542 SRETPVTPAVPPVKNPSSHKKTSKTIELKEAPATLPPSPTKSPKIPSSKKAPRT--SAPK 1599

Qy    802 E-PGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKI 860

Db   1600 EFPASPSIK-PVTT--SLAQTAPPSLQKAPSTTIPKENLAAPAV----LPVSSKSPAAPA 1652

Qy    861 TATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATN 920

Db   1653 RASASLSPATAAPQTAPKEATTIPSCKKAAATETPIETSTAPSLEGAPKETSE----TSV 1708

Qy    921 SKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAML 980

Db   1709 SKVLMSSP------PKKASSSKRASTLP-------------ATTLPSLKEASVLS---- 1744

Qy    981 QTTTRPNQTPNSKLVEVNPKSEDAGGAEGETP 1012

Db   1745 -----PTATSSGKDSHISPVS-DACSTGTTTP 1770
```

```
RESULT   8
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episia
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precurso
```

```
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218; S51
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene enc
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124; GB:
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl-en
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129; GB:
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl-en
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burchel
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epitheli
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated seque
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky, J.
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may genera
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <WR2
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associated
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu, H
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut agglu
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokine r
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 12
 partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorp
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice f
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predic
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

  Query Match              9.5%;  Score 672;  DB 1;  Length 1344;
  Best Local Similarity   27.8%;  Pred. No. 9e-22;
  Matches  310;  Conservative   78;  Mismatches  472;  Indels  256;  Gaps   56;

Qy    183 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 239
          |||   ||      |  |     :|| | || ||| :  |:  :: |: :: || ||:  :
Db     15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy    240 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 272
          |    :||:|    |    ||               | |     ||||          :
Db     75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy    273 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 312
          | | :       : ||| |   ||| :|        ||| |   ||| :|  |       |
Db    135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy    313 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 358
          | |   ||| :|  ||   | ||| |   |  | |           :||| :   |
Db    191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy    359 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 395
            || |   |  | |           ||| :  |        || |   |  |||
Db    248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy    396 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 448
            || |   |  | |       || |   |  | |     |:   |: || :|  ||
Db    308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy    449 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 497
          | ||| |   |:| :|  |       || |   | | |       || |   |  | | |
Db    368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy    498 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 553
             || |   |  |||      |:  :| |       |       || |   |  | | |
Db    425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy    554 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 593
             || |   |  | |                ||| | | |    :||:|   |  | | |
Db    485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy    594 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 640
             || |   |  | |    |     :|| |   |  | |           :|||
Db    545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy    641 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 694
          ||      ||    :|||  | |     |:    || |   |  | |      :|| |
```

```
Db    601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy    695 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 738

Db    656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy    739 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 781

Db    715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy    782 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 830

Db    773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy    831 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 881

Db    833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy    882 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 939

Db    889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy    940 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 991

Db    947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy    992 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 1051

Db   1004 AHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy   1052 ICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1087

Db   1050 ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 1085
```

RESULT 9
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of chick
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1

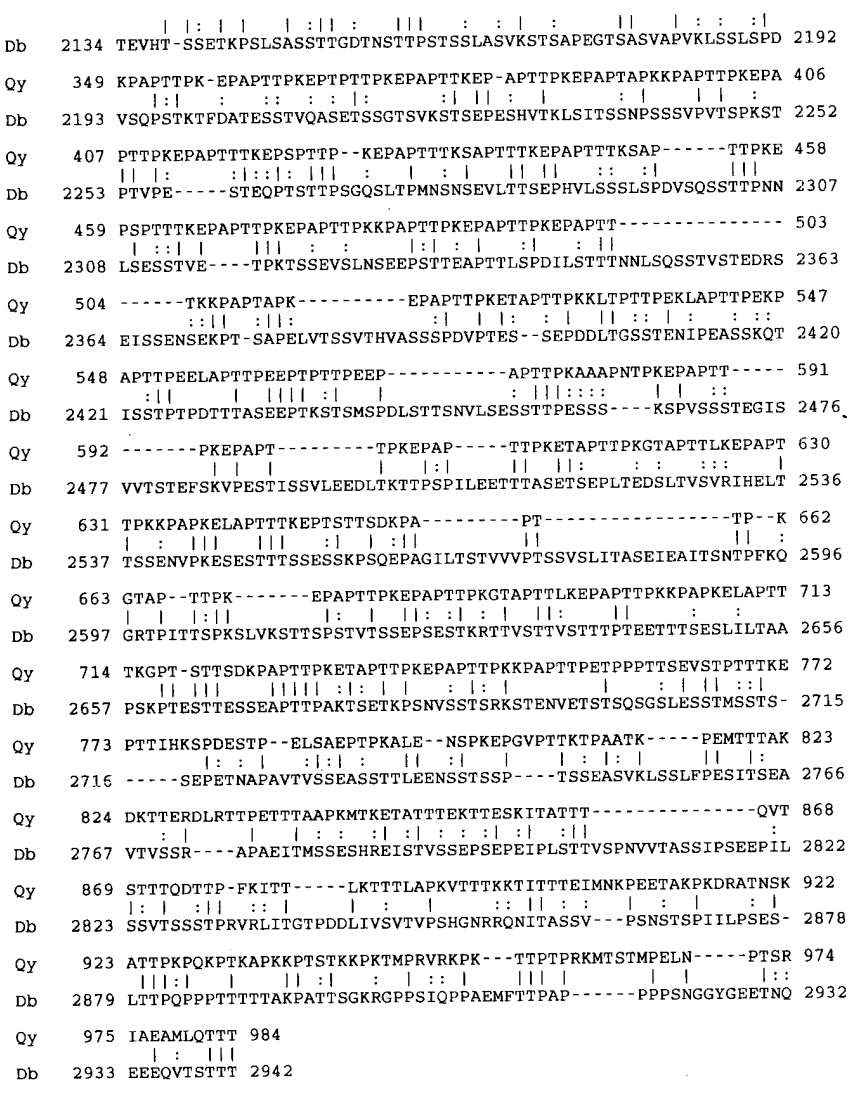
Query Match 9.5%; Score 671; DB 2; Length 1151;
Best Local Similarity 25.9%; Pred. No. 8.6e-22;
Matches 284; Conservative 100; Mismatches 419; Indels 292; Gaps 55;

```
Qy    120 TPKP-------PVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPS 170

Db    113 TPPPSQGPAGTPPPSQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP- 170

Qy    171 LPPNSDTSKETSLTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDL 228

Db    171 -------ASAASPTVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EV 219

Qy    229 APTSKVLAKPTPKAETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPT 283

Db    220 KPAAAAVAKEAKAVTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--T 277

Qy    284 TPKEPAPTT-----------TKSAPTTPKEPAPTTTK---------------------- 309

Db    278 AEAKPVPATASLMATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEA 337

Qy    310 EPAPTTPKEPAPTTTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPK 356

Db    338 KPAPPPPPQQLPKAAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP- 396

Qy    357 EPAPTTPKEPTPTT----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKE 412

Db    397 -TAGAVPKASTGTTPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQ 448

Qy    413 P---APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKE 466

Db    449 PVTKAATTTNATPPPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAAT 508

Qy    467 P----APTTPKEPAPTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAP 512

Db    509 PGVPNAATDPQKPPPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIP 568

Qy    513 K-------EPAPTTPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA------ 556

Db    569 KPVPLMALTPQPVTAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPPGLPRALAAAKLLG 627

Qy    557 -PTTP----------EEPTPTTP----EEPAPTTP---------KAAAPNTPKE------ 586

Db    628 LPSSPVASAMHAKVTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNG 687

Qy    587 ----PAPTTPKEPAPTTPKEPAPTTPKETAPT-------------TPKGTAPT------- 622

Db    688 TLMAPMGAANTQMAPIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGA 747

Qy    623 -------------TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTP 661

Db    748 GGATQMSPMGAANTQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSP 807

Qy    662 KG--TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP 717

Db    808 MGAATPPQTPSVGAATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPP 859

Qy    718 ----TSTTSDKP-APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP------- 758

Db    860 PMAATNTTQPPPMAASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPT 919

Qy    759 PTTSEVSTPTTTKEPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKP 816

Db    920 PPPPIPPSPTAQTSPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---P 975

Qy    817 EMTTTAKDKTTERDLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDT 875

Db    976 IGVTEASPSADGARLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EA 1029

Qy    876 TPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP 935

Db   1030 AP---------TKAKRSSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA--- 1077

Qy    936 KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLV 995

Db   1078 ------------PAVGDGQQQMTPGAAQSVPP-------VTEAAVQ-------------- 1104

Qy    996 EVNPKSEDAGGAEGE 1010

Db   1105 EAAAAAAAAAGAERE 1119
```

RESULT 10
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA

```
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3

  Query Match             9.4%;  Score 662;  DB 2;  Length 1229;
  Best Local Similarity   27.1%;  Pred. No. 2.2e-21;
  Matches  314;  Conservative   91;  Mismatches  367;  Indels  388;  Gaps   54;

Qy      73 RCFESFERGRECDCDAQCKKYDKCCPDYESFCAEVKDNKKNRTKKKPTPKPPVVDEAGSG 132
           | |:  :  ::  |   |    || ||                    | |
Db      34 RDFKIVKHIKKNTCTCSC----KCVPD-------------------APSNP-------- 61

Qy     133 LDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETT-- 190
           | |:|  :|    | |   |                  | |  |   :|   ||  :|
Db      62 -----FDVSTTISSINNDN-VDIGPS---------GDSNPTGSSWFQEIEATVGGQTVKS 106

Qy     191 ---------VETKETTTTNKQTS---TDGKEKTTSA---------KETQSIEKTS----A 225
                   || | ||:|:  |:   | ||: ||             :::|:   |    :
Db     107 EHNIDSSVEVEKKVTTSTDASTTNAPTTGKDSTTPEIITGIVVINSKSESVTDMSTTRFS 166

Qy     226 KDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTP 285
           |:||:::|  |    ||     : |: :: :|    | ||  | |   ||   | ||
Db     167 TTLSPTTELLTSP----ETLVSTDSSTSTEQTSPDNTTEIAS--PMETNTTT--EATTTS 218

Qy     286 KEPAPTTTKS-----------APTTPKEPAPTTTKEP---APTTPKE---PAPTTTKEPA 328
           ||: :|  |            : ||      |||:||   | :| |:    || ||:||
Db     219 VEPSVSTLASEDETTVTAIAESTTTVIAEVSTTTEEPTTTAESTTKKSTTKAPATTEEPT 278

Qy     329 PTTTKSAPTTPKEPAPTTPKKPA---PTTP----KEPAPTTPKEPT---PTTPKEPAPTT 378
           ||||:   ||  | : ||  : :   ||||    |   | | |  |   | |   |    |
Db     279 PTTTEEVTTTEAETSTTTSSETSTEKPTTPLIDNKIAGPATGKPETTHFPVTGTTPNFDT 338

Qy     379 KEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-----PSPTTPKEPAPTT 433
           |  |          | | |   :          |   |||      |  | |     :|
Db     339 ATETPFVAKSEDKMTLSKTAATETTQQTTEVTDGPEKETTKNVSIEIPITTVPLVETTST 398

Qy     434 TKSAP-------TTTKEPAPT----TTKSAPTTPKEPSPTTTKE---PAPT-------TP 472
           | :|         || |    |    :|:|| |       ||||    | ||        ||
Db     399 TSTASKESDGFHTTLKLKVTTADSDSTESATTVKPFNEETTTKSHVVPKPTKKGTVKVTP 458

Qy     473 K------EPAPTTPKKPAP-------------------TTPKE---------------P 491
           |      ||   | | | |                    :  ||
Db     459 KLELSFDEPTEIT-KAPHPGKLLEKKTYHFVLSDNFARYSEAKENDDYNHLDYNHYREAK 517

Qy     492 APTTPKEPAP----TTTKKPAPTA--PKEPAPTT---PKETAPTTPKKLTPTTPEKLA-- 540
           ||| :| :     |||::|| |    |   |||    |  || :|   |  || : :
Db     518 EPTTTEESSTTEEVTTTEEPANTGNPPTTENPTTTEQPTSTAESTTTALPFTTEQTVTTE 577

Qy     541 -PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT 599
           ||| ||   |      ||| :|   |   |   || ||   :| :|  |||  ||  ||
Db     578 EPTTAEKSTATQ----KPTTTQESVST---EKTSTTKKA---STTEE--PTTTDEPTTTT 625

Qy     600 PKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAP- 658
             : || | | |     :  ||  |   ||           |||:|||:|     |
Db     626 ---ESSTTGKATTPELSTTSEETTTTELKITTE--------GSTTTEEPTTTAIFAEAST 674

Qy     659 ---TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP------------- 701
           || : |  ||    |  |: ||     |       |::     :||
Db     675 GIITTDEETTSTTSTTPEITSTKE--IVTESAITQTSVSVVESSTPRQLPERWKAIVNKF 732

Qy     702 ---------KKPAPKELAPTTTKGP---------------TSTTSDKPAPTTPKETAPT 736
                   ||   ||   |:| |                 |:|  :|   |||  |  :
Db     733 KHNLEVLKEKKRLLKEKESTSTTGSDSSETTTVVAENIDEVTTTEKEKVVQTTPITTEKS 792

Qy     737 TPKEPAPTTPKKPAPTTPE--TPPPTTSE-VSTPTTTKEPTTIHKSPDESTPELSAEPTP 793
           | :|   ||       || :  |  |||||  :| ||| ||:|      |||
Db     793 TTQEETTTTTTTTEKTTSKTTTEKPTTSESATTETTTSEPST-----TEST--------- 838
```

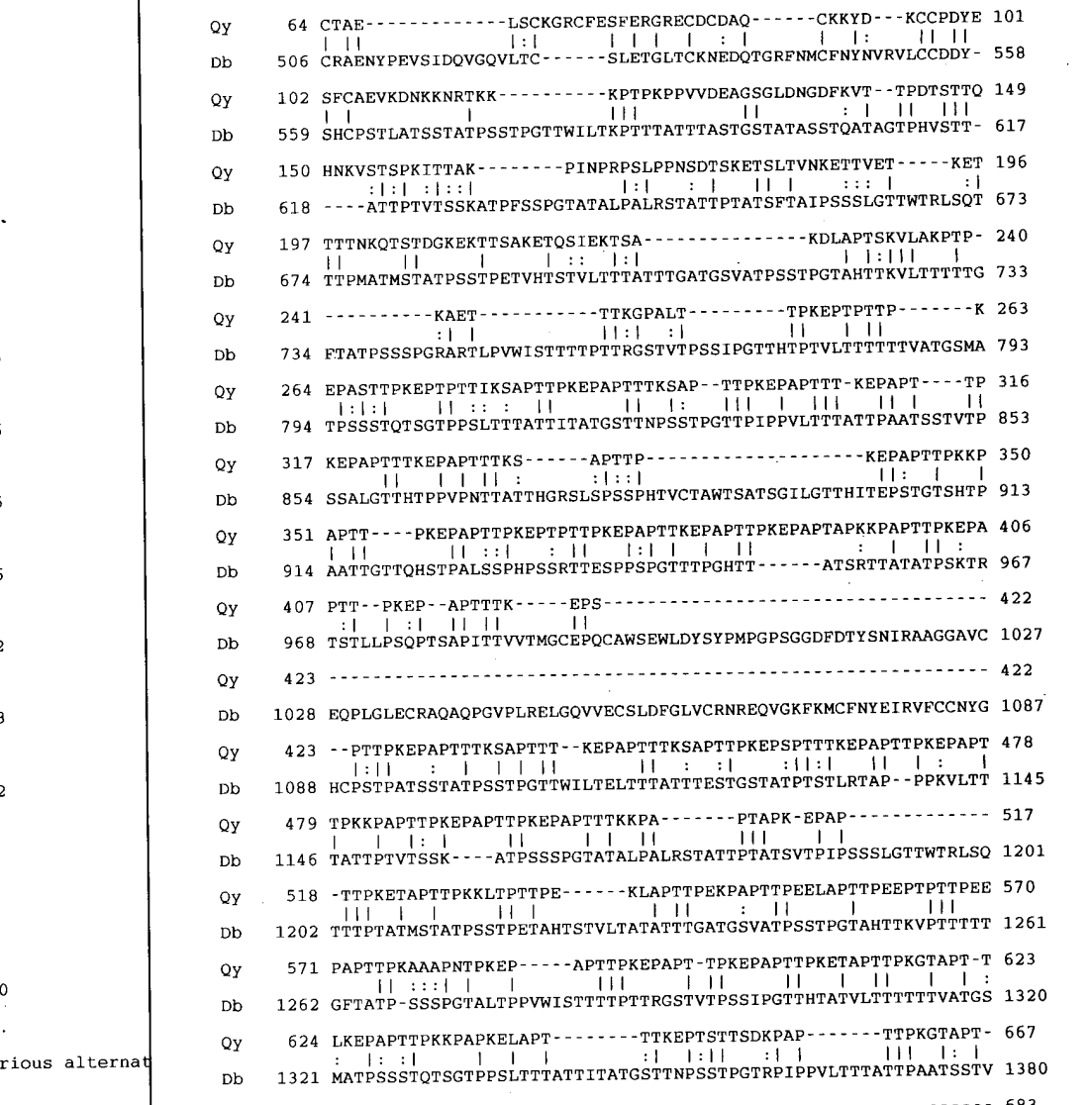
```
Qy     794 KALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA------------ 840
                   |  | :||  | :| |:  ||  :   | ||||:
Db     839 ------------TVDTSSATTEESSTAAETTTTSAE---TSETTTSESAAFITGESPEN 882

Qy     841 -------------------------APKMTKETATTTEKTTESKITATTTQ-VTST--- 870
                                    ||  | |    | |:    : |||: :|:|
Db     883 TALQSSSQKSEENESSAEKPGARRDFVPKKHKTTVKPAETTSAVAASTTTTEPITTTEKS 942

Qy     871 TTQDTTPFKITTLKTTTLAPKVT---TTKKTITTTEIMNKPEETAKPKDRATNSKATTPK 927
           || :||| : |||   |    ||     : || | |:::|           |:: : ||
Db     943 TTLETTPIEATTLNEVTGPAFVTGAPVDETTINTLELLSK----------INNTQISQPK 992

Qy     928 PQ----------------KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELN 970
           |                 ||||  ||               ||     :|:|   ||
Db     993 PTDISKTDALSSLISGLIGSFTKAPMAPTI------------HTTTDAAFVTATEASLN 1039

Qy     971 ------------PTSRIAEA 978
                       ||  |  |
Db    1040 DGSDKKIIDEAQPTDEIRRA 1059


RESULT  11
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/
 3504/1

  Query Match             9.4%;  Score 661;  DB 2;  Length 3507;
  Best Local Similarity   23.0%;  Pred. No. 6.6e-21;
  Matches  278;  Conservative  173;  Mismatches  459;  Indels  300;  Gaps   53;

Qy      27 LSSCAGRCGEGYSRD-ATC-------NCDYNCQHYMECCPDFKRVCTAE---------- 67
           : ||  :|  ||: | |||          |        |  :: |:|  |
Db    1781 MGSCGCKCMAGYTGDGATCIKIEEEPKSDKTA-----CTDEWSRLCELEKKQCTVDEEEV 1835

Qy      68 -------------------LSCKGRCFESFERGRECDC-----DAQ------------- 89
                             |   | | :   :  :|    |:
Db    1836 PQCGACLPGHHPINGTCQSLQISGLCAQKNDCNKHAECIDIHPDSHFCSCPDGFIGDGMI 1895

Qy      90 CKKYDKC-----CPDYESFC--------------AEVKDNKKNRTKKKPTPKPPVVDEA 129
           |   |:|     | |  : |                :| :       ||:|  :   :||
Db    1896 CDDVDECNNAGMCDDENTKCENTIGSFNCVCLEGFKKVDEKCVVDEKKQPNREKIEIDEE 1955

Qy     130 GSGLDN-GDFKVTT----PDTSTTQHNKVSTSPKITTA-------KPINPRPSLPPN--- 174
           |   | |  | ||      || |     :  | :||:        | :    | | |
Db    1956 NSSSSNSGQEKPTTKGIVSSTSATSSESTTAEPHVTTSISSTTSTKDMTSSKS-PENVTM 2014

Qy     175 SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAP--- 230
           |  | | | | : :| ||  ||  ::| :: :|::    :: |  |: | ::| |   |
Db    2015 SSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEVITESSVKSTTPKEE 2074

Qy     231 -TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKS-APTTPKEP 288
            :|::  | : |:   |:     ::|  |: || :   || |:    | : | || |   |
Db    2075 SSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKSTVLSSEAPVTSTSP 2133

Qy     289 APTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPK 348
```



RESULT 12
T45025
mucin MUC5B, tracheobronchial [imported] - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 21-Jan-2000 #sequence_revision 21-Jan-2000 #text_change 21-Jul-2000
C;Accession: T45025
R;Desseyn, J.L.; Guyonnet-Duperat, V.; Porchet, N.; Aubert, J.P.; Laine, A.
J. Biol. Chem. 272, 3168-3178, 1997
A;Title: Human mucin gene MUC5B, the 10.7 kb large central exon encodes various alternat
A;Reference number: Z22899; MUID:97166151
A;Accession: T45025
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA

A;Residues: 1-3570 <DES>
A;Cross-references: EMBL:Z72496; NID:g1834502; PIDN:CAA96577.1; PID:g1834503
A;Experimental source: placenta
C;Genetics:
A;Gene: MUC5B

Query Match 9.2%; Score 646.5; DB 2; Length 3570;
Best Local Similarity 22.5%; Pred. No. 2.8e-20;
Matches 359; Conservative 107; Mismatches 536; Indels 591; Gaps 65;



```
Db   1381 TPSSALGTTHTPPVPNTTATTHGRSLSPSSPHTVRTAWTSATSGTLGTTHITEPSTGTSH 1440

Qy    684 TPKGTAPTT------LKEPAP---TTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETA 734
          ||  |  ||      |  | |   ||   |:|     ||| | |: ||   |  || :|
Db   1441 TPAATTGTTQHSTPALSSPHPSSRTTESPPSP----GTTTPGHTTATSRTTATATPSKTR 1496

Qy    735 PTT--PKEP---------------------------------------------------- 741
          :|  |  |
Db   1497 TSTLLPSSPTSAPITTVVTMGCEPQCAWSEWLDYSYPMPGPSGGDFDTYSNIRAAGGAVC 1556

Qy    742 ---------------------------------------------------------- 741

Db   1557 EQPLGLECRAQAQPGVPLRELGQVVECSLDFGLVCRNREQVGKFKMCFNYEIRVFCCNYG 1616

Qy    742 -APTTPKKPAPTTPETPPPTT---SEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALE 797
           |:||   :  || : | ||   :| :|  ||   |     |  |||     | || |
Db   1617 HCPSTPATSSTATPSSTPGTTWILTEQTTAATTTATTGSTAIP-SSTP--GTAPPPKVLT 1673

Qy    798 NSPKEPGVPTTKTPAATKPEMTTT------AKDKTTERDLRTTPETT------------- 838
          :   |   ::|  ::: |   ||         |:|       | :|
Db   1674 SQATTPTATSSKATSSSSPRTATTLPVLTSTATKSTATSFTPIPSSTLGTTGTSQNRPPH 1733

Qy    839 ------TAAPKMTKETA-TTTEKTTESKITATTTQVTST-TTQDTTP------FKITTLK 884
                |  |  | ||  |:|  ||:    ||||: ||: :|  :||       :::||
Db   1734 PMATMSTIHPSSTPETTHTSTVLTTK----ATTTRATSSMSTPSSTPGTTWILTELTTAA 1789

Qy    885 TTTLAPKVTTTKKTITTTEIMNKPEETA----------KPKDRAT--NSKATTPKPQKP 931
          |||  |    |   || |: :|  ||              |||    |  |     |
Db   1790 TTTAALPHGTPSSTPGTTWILTEPSTTATVTVPTGSTATASSTRATAGTLKVLTSTATTP 1849

Qy    932 TKAPKKPTSTKKPKT---MPRVRKPKTTPTPRKMTSTMPELNPTSRIAEA--MLQTTTRP 986
          |    : | :  | |   :| :|   ||||   :|:      |:| :  |   |  || |
Db   1850 TVISSRATPSSSPGTATALPALRSTATTPTATSVTAI-----PSSSLGTAWTRLSQTTTP 1904

Qy    987 NQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVP---NQGIIIN 1043
            | ::      |::        |       |  |  ||   :  :||     |
Db   1905 TATMSTATPSSTPETVHTSTVLTTTATTTRTGSVATPSSTPGTAHTTKVPTTTTGFTAT 1964

Qy   1044 PMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPSPARRI---------- 1092
          |  |  | :   ||   ||    |   ||    ::| | |
Db   1965 PSSSPGTALT--PPVWISTT----TTPTTRGS---TVTPSSIPGTTHTATVLTTTTTTVA 2015

Qy   1093 ----------TEVWGIPSPIDTVFTRCNCEGKT 1115
                    |:  | |  : |  |     | |
Db   2016 TGSMATPSSSTQTSGTPPSLTTTATTITATGST 2048


RESULT  13
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31


  Query Match             9.0%;  Score 635.5;  DB 2;  Length 7962;
  Best Local Similarity   22.8%;  Pred. No. 1.8e-19;
  Matches  306;  Conservative  112;  Mismatches  475;  Indels  451;  Gaps   63;

Qy     76 ESFERG-----RECDCDAQCKKYDKCCPDYE--SFCAEVKDNKKNR-------TKKKPTP 121

          | || |     ||   |   :::::   : |      || :    |        || | |
Db   5783 EDFEEGQEYYEREEGYDEGEEEWEEAYQEREVIQVQKEVYEESHERKVPAKVPEKKAPPP 5842

Qy    122 ----KPPV---VDEAGSGLDNGDFKVT-----------TPDTSTTQHNKVSTS-PKITT 161
              | ||   :::    ::    :||                |  :  |   :    | : |
Db   5843 PKVIKKPVIEKIEKTSRRMEEEKVQVTKVPEVSKKIVPQKPSRTPVQEEVIEVKVPAVHT 5902

Qy    162 AKPINPRPSLPPNSDTSKETSLT---VNKETTVETKETTTTNKQTSTDGKEKTTSAKETQ 218
            | :     :   | | :| |:|   | ||   | |    :|:                |  :
Db   5903 KKMVISEEKMFFASHTEEEVSVTVPEVQKEIVTEEKIHVAVSKRVE-------PPPKVPE 5955

Qy    219 SIEKTSAKDLAPT--SKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPAS-----TTPK 271
            || : :::||    | :  | ||     | | :   |:|'|   ||||:       ||
Db   5956 LPEKPAPEEVAPVPIPKKVEPPAPKVPEVPKKP-VPEEKKPVPVPKKEPAAPPKVPEVPK 6014

Qy    272 EPTPTTIKSAPTTPKEPAP-----------TTTKSAPTTPKEPAPTTTKEPA-PTTPKE 318
            :| |       |   |: ||           | |    | :| :|     || |  ||:
Db   6015 KPVPEEKIPVPVAKKKEAPPAKVPEVQKGVVTEEKITIVTQREESP----PPAVPEIPKK 6070

Qy    319 PAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP-KEPAPTTPKEPTPTT------- 370
            |   | | |   :  |  || ||   |||| |      | |   | | |
Db   6071 KVPEERK-PVPRKEEEVPPPPKVPA--LPKKPVPEEKVAVPVPVAKKAPPPRAEVSKKTV 6127

Qy    371 ----------------------------------------------------PKEPAPT 377
                                                               :| |
Db   6128 VEEKRFVAEEKLSFAVPQRVEVTRHEVSAEEEWSYSEEEEGVSISVYREEEREEEEEAEV 6187

Qy    378 TK-------------------------EPAPTT----------PKEPA----------PT 392
            |:                        |||  |          || ||          |
Db   6188 TEYEVMEEPEEYVVEEKLHIISKRVEAEPAEVTERQEKKIVLKPKIPAKIEEPPPAKVPE 6247

Qy    393 APKK-------PAPTTPKE--PAPTTPKEP-APTTTKEPSPTTPKE----PAPTTTKSAP 438
            ||||       |||   ||  | |  |:|| |   |:  |   |     ||  | |
Db   6248 APKKIVPEKKVPAPVPKKEKVPPPKVPEEPKKPVPEKKVPPKVIKMEEPLPAKVTEKHMQ 6307

Qy    439 TTTKEP--APTTTKSAP---TTPKEPSPTTTKE-------------PAPTT--------- 471
            | :|      | | ||     |:||     :|              ||  |
Db   6308 ITQEEKVLVAVTKKEAPPKARVPEEPKRAVPEEKVLKLKPKREEEPPAKVTEFRKRVVKE 6367

Qy    472 -------PK-EPAP---------------------------------------------- 477
                  || || |
Db   6368 EKVSIEAPKREPQPIKEVTIMEEKERAYTLEEEAVSVQREEEYEEYEEYEDYKEFEEYEPT 6427

Qy    478 ------------------------TTPKKPAPT--TPKEPAPTTPKEPAPTTTKKPAPTA 511
                                   | |:|| |    |:|| || || |     ||  |
Db   6428 EEYDQYEEYEEREYERYEEHEEYITEPEKPIPVKPVPEEPVPTKPKAPPAKVLKKAVP-E 6486

Qy    512 PKEPAPTTPKETAPTTPKKLTPTTPEK-------LAPTTPEKPAPTTPEELAPTTPEEPT 564
            | | |  ||:  |  ||   |  |:|       :: |  ||   | |   |   | :|
Db   6487 EKVPVP-IPKKLKPPPPK--VPEEPKKVPEEKIHISITKREKEQVTEP---AAKVPMKPK 6540

Qy    565 PTTPEEPAPTTPKAAAP--NTPKEPAPTTPKEPA------------------------- 596
              ||  |   |  ||    |: |    |:| |
Db   6541 RVVAEEKVPVPRKEVAPPVRVPEVPKELEPEEVAFEEEVVTHVEEYLVEEEEEYIHEEEE 6600

Qy    597 PTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTT---- 652
            | :|  |  | :  |  |:   |   |:| |   || ||    |   |:|
Db   6601 FITEEEVVPVIPVK-VPEVPRKPVPEE-KKPVPVPKKKEAPPAKVPEVPKKPEEKVPVLI 6658

Qy    653 --SDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKPAPKE 708
              :||    |    |  ||:| |   | | |   |  ||     ||  |  |||| |::
Db   6659 PKKEKP----PPAKVPEVPKKPVP-EEKVPVPVPKKVEAP-----PAKVPEVPKKPVPEK 6708

Qy    709 L----APTTTKGPTSTTSDKPAPTTPKETAPT-TPKE---PAPTTPKKPAPTTPETPPPT 760
              ||   : | :   : |    |:|  ||  ||:   | |  |||  |       |
Db   6709 KVPVPAPKKVEAPPAKVPEVPKKLIPEEKKPTPVPKKVEAPPPKVPKKREPVPVPVALPQ 6768

Qy    761 TSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTT 820
            ||          :   :|  ||    |  |:  |  |:|  :|  :     : |
Db   6769 EEEVLFEEEIVPEEEVLPEEEEVLPE-EEEVLPEEEEVLPEEEEIPPEEEEVPPEEEYVP 6827

Qy    821 TAKDKTTERDL--RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPF 878
            ::   | ::     |:    ||          :| || |:     :          |
```

```
Db   6828 EEEEFVPEEEVLPEVKPKVPVPAP-----VPEIKKKVTEKKVVIPKKEEAPPAKVPEVPK 6882

Qy    879 KITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKP 938
          |:   :   : ||    :: :   |:  :|||                 :|    : |::|
Db   6883 KVE--EKRIILPK----EEEVLPVEVTEEPEE----------------EPISEEEIPEEP 6920

Qy    939 TSTKK-----PKTMPRVRK---PKT-TPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 989
          | ::     |  :| | |   |:  || |:|: :  |    : :| |  :    :  :
Db   6921 PSIEEVEEVAPPRVPEVIKKAVPEAPTPVPKKVEA--PPAKVSKKIPEEKVPVPVQKKEA 6978

Qy    990 PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDE 1049
          | :|: ||                          |:  |:   |  || :  :
Db   6979 PPAKVPEV------------------------PKKVPEKKVL--VPKKEAV-------- 7003

Qy   1050 TNICNGKPVDGLTTLRNGTLVAFR 1073
                |  | | |     ||||
Db   7004 ------PPAKGRTVLEEKVSVAFR 7021


RESULT  14
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume inte
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>


  Query Match             9.0%;  Score 633;  DB 2;  Length 489;
  Best Local Similarity   31.2%;  Pred. No. 1.6e-20;
  Matches  149;  Conservative   33;  Mismatches  252;  Indels   44;  Gaps    7;

Qy    285 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 344
          ||:  |    :  |   | | |     |:|: |          | |   |  |   | | |
Db     39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy    345 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 401
          :|  | |   | | | :|  | |   | | | :  |:|  |   | | | :|  | |
Db     84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy    402 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 461
          | | | :|  | |   | | | :|  | |   || |           | :|  | |
Db    142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP-------------PPSPSPPPP 187

Qy    462 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 521
            | | | :|  | |   | | | :|  | |   | | | :  | |   | | | :|
Db    188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy    522 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 581

          |   |   | :|    |   : | | :|    |   : | | :|  | |   |:  |
Db    248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy    582 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 641
            :|  | |   | | | :|  | |   |   | :|    |   | | | :|  | |
Db    308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYY- 366

Qy    642 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 701
              | |   :   | |   |   | :|  | |   | | | :|    |   | | | :|
Db    367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy    702 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 759
            |      |   | |   :   | |   |   | :|  | |   | | | :|  |||
Db    423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPP 475


RESULT  15
C84672
hypothetical protein At2g27380 [imported] - Arabidopsis thaliana
C;Species: Arabidopsis thaliana (mouse-ear cress)
C;Date: 02-Feb-2001 #sequence_revision 02-Feb-2001 #text_change 02-Feb-2001
C;Accession: C84672
R;Lin, X.; Kaul, S.; Rounsley, S.D.; Shea, T.P.; Benito, M.I.; Town, C.D.; Fujii, C.Y
M.; Koo, H.; Moffat, K.S.; Cronin, L.A.; Shen, M.; VanAken, S.E.; Umayam, L.; Tallon,
euss, D.; Nierman, W.C.; White, O.; Eisen, J.A.; Salzberg, S.L.; Fraser, C.M.; Venter
Nature 402, 761-768, 1999
A;Title: Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana.
A;Reference number: A84420; MUID:20083487
A;Accession: C84672
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-761 <STO>
A;Cross-references: GB:AE002093; NID:g5306260; PIDN:AAD41992.1; GSPDB:GN00139
C;Genetics:
A;Gene: At2g27380
A;Map position: 2


  Query Match             8.9%;  Score 632;  DB 2;  Length 761;
  Best Local Similarity   31.1%;  Pred. No. 2.7e-20;
  Matches  217;  Conservative   44;  Mismatches  354;  Indels   82;  Gaps   33;

Qy    238 PTPKAETTTKGPALTTP---KEPTPT-TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTT 293
          | |  :  |  | :  |   | |||| :|       | ||||       | :  || |
Db     69 PPPIQKPPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPT 128

Qy    294 KSAPTTPKEPAPTTTKEPAP--TTPKEPAPTTTKEPAPTTTKSAPTTP----KEPAPT-- 345
          | |  |    |   | | |  : | :| |    :  || | | |   |    | | ||
Db    129 YSPPIYP----PPIQKPPTPSYSPPVKPPPV---QMPPTPTYSPPIKPPPVHKPPTPTYS 181

Qy    346 TPKKP---APTTPKEPAPTTP----KEPTP--TTPKEPAPTTKEPAPT-TPKEPAPTAPK 395
          | ||    | ||    |  |    | |||  : | :| |  | | || :|    |   |
Db    182 PPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHK 241

Qy    396 KPAPT-TPKEPAPTTPKEPAP---TTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 451
          | |  :|    |   | | |      | |   ||  |  :     |   | | ||  :
Db    242 PPTPIYSPPIKPPPVHKPPTPIYSPPVKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPP 301

Qy    452 APTTPKEPSPTTTKEP---APTTPKEPAPT--TPKKPAPTTPKEPAPTTPKEPAPTTTKK 506
          : | :  || |  |    |   | | ||   | || |  |  | : | :| |   |
Db    302 VKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPVKP-PPVHKP 360

Qy    507 PAP--TAPKEPAPT----TPKETAPTTP----KKLTPTTPEKLAPTTPEKPAPTT---PE 553
          | |  : | :| |      ||  :  |   |  |||    :  |   :||   |   |
Db    361 PTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPI 420

Qy    554 ELAPTTPEEPTPTTPEEPAPT----TPKAAAP-NTPKEPAPTTPKEPAPTTPKEPAPTTP 608
          :| |  |  |  : | :| |      ||  : |   |    | ||    |  |   | ||
Db    421 KLPPVKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTP 480

Qy    609 KETAPTTPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 666
          : |  |    |   | | ||   | || | :  ||    || :  || | |
Db    481 TYSPPVQP----PPVQKPPTPTYSPPVKPPPIQKPPT----PTYSPPIKPPPVKPPTPTY 532
```

```
Qy     667 TTPKEPAPTTPKEPAPT-TPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKGPTSTTSD 723
           : | :| |   | | || :|    |   | | ||   | || |   || |  |
Db     533 SPPIKP-PPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPP 591

Qy     724 KPAPTTPKETAPTTPKEPAPTTPKKPAPT-TPETPPPTTSEVSTPTTTK--EPTTIHKSP 780
           | ||  : |  |    |   | | || :|   ||   :  ||| :   :|  :|| |
Db     592 VHKPPTPTYSPPIKP----PPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPP 647

Qy     781 DESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA 840
           ||  |  |  ::  |     |  | |    |  | |         ::  |  | :
Db     648 ---TPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPP-TPTYSPPVKPPPVQVPPTPTYS 703

Qy     841 APKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP 877
           |             |    |     ||  | |  : |
Db     704 PPVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPP 740
```

Search completed: April 26, 2002, 16:33:15
Job time: 790 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:33:14 ; Search time 62.75 Seconds
(without alignments)
766.017 Million cell updates/sec

Title: AA8
Perfect score: 7064
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1311

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1180	16.7	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	950	13.4	1664	1	SLP1_CLOTM	Q06852 clostridium
3	798.5	11.3	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	9.2	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	556.5	7.9	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	7.9	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	7.8	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	7.5	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	7.5	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	7.5	865	1	CPN_DROME	Q02910 drosophila
11	518.5	7.3	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	7.2	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	7.1	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	7.1	1970	1	RPB1_MOUSE	P08775 mus musculu
15	497	7.0	831	1	NFH_RAT	P16884 rattus norv
16	493.5	7.0	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	6.9	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488	6.9	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	6.7	267	1	EXTN_MAIZE	P14918 zea mays (m
20	472	6.7	2142	1	BAT2_HUMAN	P48634 homo sapien
21	471.5	6.7	5376	1	ZAN_MOUSE	O88799 mus musculu
22	468.5	6.6	634	1	HWP1_CANAL	P46593 candida alb
23	467	6.6	817	1	VRP1_YEAST	P37370 saccharomyc
24	455	6.4	1161	1	YJ9P_YEAST	P47179 saccharomyc
25	454	6.4	797	1	VGLX_HSVEB	P28968 equine herp
26	448.5	6.3	670	1	VG50_HSV11	Q00130 ictalurid h
27	442.5	6.3	751	1	FP1_MYTGA	Q27409 mytilus gal
28	439.5	6.2	1083	1	T2D3_HUMAN	O00268 homo sapien
29	432.5	6.1	439	1	XP2_XENLA	P17437 xenopus lae
30	432	6.1	3164	1	TEGU_HSV11	P10220 herpes simp
31	430.5	6.1	3421	1	TEGU_HSVEB	P28955 equine herp
32	426	6.0	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
33	424.5	6.0	1125	1	MAP4_MOUSE	P27546 mus musculu
34	424	6.0	307	1	SGS3_DROME	P02840 drosophila
35	424	6.0	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
36	422.5	6.0	1794	1	YAV1_SCHPO	Q10172 schizosacch
37	422	6.0	2774	1	MAPA_RAT	P34926 rattus norv
38	421	6.0	1229	1	N121_HUMAN	Q9y2n3 homo sapien
39	419.5	5.9	2476	1	ZAN_PIG	Q28983 sus scrofa
40	419	5.9	1185	1	DRPL_HUMAN	P54259 homo sapien
41	418.5	5.9	1411	1	TCOF_HUMAN	Q13428 homo sapien
42	411.5	5.8	907	1	VGP3_EBV	P03200 epstein-bar
43	410	5.8	1183	1	DRPL_RAT	P54258 rattus norv
44	408.5	5.8	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
45	407.5	5.8	3256	1	KI67_HUMAN	P46013 homo sapien

ALIGNMENTS

```
RESULT   1
MUC2_HUMAN
ID   MUC2_HUMAN      STANDARD;      PRT;  5179 AA.
AC   Q02817; Q14878;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN   MUC2 OR SMUC.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Intestine;
RX   MEDLINE=94132002; PubMed=8300571;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT   "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT   Identification of the amino terminus and overall sequence similarity
RT   to prepro-von Willebrand factor.";
RL   J. Biol. Chem. 269:2440-2446(1994).
RN   [2]
RP   SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC   TISSUE=Colon;
RX   MEDLINE=93016075; PubMed=1400449;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA   Kim Y.S.;
RT   "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT   both upstream and downstream of its central repetitive region.";
RL   J. Biol. Chem. 267:21375-21383(1992).
RN   [3]
RP   SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX   MEDLINE=91358717; PubMed=1885763;
RA   Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA   Petersen G.M., Kim Y.S.;
RT   "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT   and polymorphism.";
RL   J. Clin. Invest. 88:1005-1013(1991).
CC   -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC       OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC       PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC       AGENTS AT MUCOSAL SURFACES.
CC   -!- SUBUNIT: MULTIMERIC.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC       BRONCHUS, CERVIX AND GALL BLADDER.
CC   -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC       INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC       VARIES AMONG DIFFERENT ALLELES.
CC   -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC       OF SILKWORM HEMOCYTIN.
CC   -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC   ---------------------------------------------------------------------
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;

  Query Match             16.7%;  Score 1180;  DB 1;  Length 5179;
  Best Local Similarity   30.4%;  Pred. No. 4e-41;
  Matches  389;  Conservative   68;  Mismatches  512;  Indels  310;  Gaps   48;

Qy       4 KTLPIYLLLLLSVFVIQQVSSQDLSSC--AGRCG-----EGYSRDATCNCDYNCQHYMEC 56
           |  |||                :||  |  | :||       |   |:   :  |:  : |
Db    1169 KDRPIY-------------EEDLKKCVTADKCGCYVEDTHYPPGASVPTEETCKSCV-C 1213

Qy      57 CPDFKRVCTAELSCKGRCFESFERG-----RECDCDAQCKKYDKCC---------PDYES 102
             : ||  |   :|:     : |       |  :   :|:   |           : :
Db    1214 TNSSQVVCRPE---EGKILNQTQDGAFCYWEICGPNGTVEKHFNICSITTRPSTLTTFTT 1270

Qy     103 FCAEVKDNKKNRTKKKPTPKPPVV----------------DEAGSGLDNGDFK------ 139
                       |    ||    |                |   || |:|| :
Db    1271 ITLPTTPTSFTTTTTTTTPTSSTVLSTTPKLCCLWSDWINEDHPSSGSDDGDREPFDGVC 1330

Qy     140 ----------VTTPDTSTTQH-NKVSTSPK---------------------------- 158
                    |  |  |  ||  ||
Db    1331 GAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIRVNCCW 1390

Qy     159 -----ITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTS 213
                |||  |    || || : |:   : | :  ||     ||||   |:|     ||:
Db    1391 PMDKCITTPSPPTTPSPPPTTTTTLPPTTTPSPPTT-----TTTTPPPTTTPSPPITTT 1445

Qy     214 AKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPT-----PTTPKEPAST 268
                          ||:     |:|   |||  |  |||  ||      |||   | :|
Db    1446 TTP-----------LPTT----TPSPPISTTTTPPPTTTPSPPTTTPSPPTTTPSPPTT 1489

Qy     269 TPKEPTPTTIKSAP-TTPKEP-------APTTTKSAP-TTPKEPAPTTTKEPAPTTPKEP 319
           |   | |||  | | |||  |         |||| | | ||   | ||||  |  |||  |
Db    1490 TTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITP 1549

Qy     320 -------APTTTKEPAPTTTKSAP--TTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTT 370
                   ||||  | |||| : |  |||  |  |||  |  ||   |  |||  || ||
Db    1550 PTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTPSPPTITTTTPPPTTTPSPPTTTT 1609

Qy     371 PKEPAPTTKEPAPTTPKEP-------APTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSP 423
           |  ||  |  |||  |         ||    | |||   | |||   |  |||  |
Db    1610 TTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPI 1669

Qy     424 TTPKEPAPTTTKSAPTTTKEPAPTTTKSAP---TTPKEPSPTTTKEPAPTTPKEPAPTTP 480
           ||   | |||| |:| ||    ||||  : |   |||  |  |||  |: ||   | |||
Db    1670 TTTTTPPPTTTPSSPITTTPSPPTTTMTTPSPTTTPSSPI-TTTTTPSSTTTPSPPPTTM 1728

Qy     481 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTP-----KETAPTTP------- 528
             |:|||   |  ||     ||||  |  | |  |: | |        | ||||
Db    1729 TTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPTFSPFSTTTPTTPCVPLCNW 1788
```

```
Qy     529 ------------------------------------KKLTP------------------- 533
                                               : |
Db    1789 TGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYPDVPIGQLGQTVVCDVSVGL 1848

Qy     534 -----------------------------TTPEKLAPTTPEKPAPTTPEELAPTTPEE 562
                                         | |  :  || | | | |   :  ||
Db    1849 ICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTMTTTTTENPTPPTTTPITTTTTVT 1908

Qy     563 PTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKEP-----APTTPKEPAPT-TPKETAPTT 615
           |||| |  :   |||          | ||  : |       ||   | || |  :|  ||
Db    1909 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT 1968

Qy     616 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAP-TTPKGTAPTTPKEPAP 674
           |  |  |    | ||  : |    :  |||  || | :    | |||  |  |     | |
Db    1969 PITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP 2028

Qy     675 TTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAP-TTPKET 733
           |  : |  |||  |  |    | ||  : |    :  |||  || | :    | |||  |
Db    2029 TGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT 2087

Qy     734 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTP 793
             |    | ||  : |  ||| |   |  :   ||| |: |||   :|  :|  ::  |||
Db    2088 TTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTT---TPITTTTTVTPTPTP 2143

Qy     794 KALENSPKEP----------GVPT-TKTPAATKPEMTTTAKDKTTERDLR---TTP--ET 837
           :  |          || |:||  |    |||      |    :   |||   |
Db    2144 TGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTT 2203

Qy     838 TTAAPKMT---KETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTT 894
           ||  |  |    :| |||  || : :| | |  | | |  ||| |||  |||: |  |
Db    2204 TTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP-TGTQTPTTTP--ITT--TTTVTPTPTP 2258

Qy     895 T-KKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP-KKPTSTKKPKTMPRVRK 952
           |  :| ||| |       |  |    | :  |||     |  |   || |: | | |
Db    2259 TGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTT 2318

Qy     953 PKTTPTPRKMTSTMPELNP---TSRIAEAMLQTTTR-PNQTPNSKLVEVNPKSEDAGGAE 1008
             ||||    :  |  |   |: :       | |: |  || :    | |     |
Db    2319 TTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTG--- 2375

Qy    1009 GETPHMLLRPHVFMPEVTP 1027
           :||     |     |||
Db    2376 TQTP--TTTPITTTTTVTP 2392


RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL        1     28       POTENTIAL.
FT   CHAIN        29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN       36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT       36    191       1.
FT   REPEAT      207    363       2.
FT   REPEAT      409    565       3.
FT   REPEAT      607    763       4.
FT   DOMAIN      771   1377       APPROXIMATE TANDEM REPEATS OF
FT                                T-P-S-D-E-P.
FT   DOMAIN     1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN     1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN     1495   1565       SLH 2.
FT   DOMAIN     1566   1625       SLH 3.
FT   DOMAIN     1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;


  Query Match             13.4%;  Score 950;  DB 1;  Length 1664;
  Best Local Similarity   31.0%;  Pred. No. 3.5e-32;
  Matches  312;  Conservative  104;  Mismatches  361;  Indels  230;  Gaps   55;

Qy     234 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 292
           |: :| | || :    |  |   ||||:          ||||:    ||  || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     293 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 348
             :  ||:|| || |    ||   ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     349 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 407
             ||  || |:   ||||:   || | : || |: ||:|| ||      ||  || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     408 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 460
           :    |:  || : ||:|: ||:|| || | |  || : || |:    |    ||   ||:
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy     461 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 515
             || : || |: ||:|| ||      ||   || |:    |:   |    ||   ||
Db     980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy     516 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 568
           |:   || |    ||::  ||      ||  ::| |:    || :  ||  :||||: ||
Db    1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy     569 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 625
           ||| ||   :  |    || |:    |:  ||   || |: ||:|  ||      ||
Db    1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy     626 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 684
           || |:   :| | :  || : ||| | | ::| ||      ||   || |:   || |:
Db    1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy     685 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 740
             ||   || |: ||::|            || | ||:| |:   || :  ||    |
Db    1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy     741 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 799
           | |:   :| |: ||| | || :    || : |||      || ||  | ||||
Db    1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305
```

```
Qy     800 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 857
Db    1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy     858 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 917
Db    1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy     918 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 962
Db    1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy     963 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 1014
Db    1454 YLRGYPDGSFRPERNITRAEAAVIF-----------AKLL----GADESYGAQSASPYSD 1498

Qy    1015 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 1040
Db    1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy    1041 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 1087
Db    1559 TIDISNPKFDD---CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy    1088 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1132
Db    1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

```
RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;
```

```
Query Match           11.3%;  Score 798.5;  DB 1;  Length 1367;
Best Local Similarity 27.9%;  Pred. No. 4.2e-26;
Matches  313;  Conservative  108;  Mismatches  523;  Indels  177;  Gaps   50;

Qy      85 DCDAQCKKYDK---CCPDYESFCAEVK-DNKKNRTKKKPTPKPPVVDE--AGSGLDNGDF 138
Db     174 DLSTGCNNYDNQGHSQTDFPGFYWNIDCDNNCGGTKSSTTTSSTSESSTTTSSTSESSTT 233

Qy     139 KVTTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETT 190
Db     234 TSSTSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTT 293

Qy     191 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKG 248
Db     294 PCTKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSA 342

Qy     249 PALTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPT 298
Db     343 PVTSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSST 402

Qy     299 TPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP 358
Db     403 TESSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTT 451

Qy     359 APTTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPK 411
Db     452 ESSSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTES 507

Qy     412 EPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEP 467
Db     508 SSAPVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSS 566

Qy     468 AP----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKE 522
Db     567 TPVTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESS 622

Qy     523 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTP 576
Db     623 SAPVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 677

Qy     577 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKP 635
Db     678 SSSAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSST 729

Qy     636 APKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPT 691
```

```
Db    730 TESSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 788

Qy    692 TLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTP 746

Db    789 SSSAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTE 848

Qy    747 KKPAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPK 801

Db    849 SSSAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSK 904

Qy    802 EPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKIT 861

Db    905 YPGSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVT 964

Qy    862 AT--TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPK 915

Db    965 TTVPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA--- 1021

Qy    916 DRATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPT 958

Db   1022 -----SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPT 1076

Qy    959 PRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRP 1018

Db   1077 P--SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGCSPKTVTTTVPC 1120

Qy   1019 HVFMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFW 1078

Db   1121 STGTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVT 1173

Qy   1079 MLSPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1115

Db   1174 TLAPSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207
```

```
RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
```

```
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23        POTENTIAL.
FT   CHAIN        24   1255        MUCIN 1.
FT   DOMAIN       24   1162        EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186        POTENTIAL.
FT   DOMAIN     1187   1255        CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960        44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151        SEA.
FT   CARBOHYD    957    957        N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975        N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029        N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055        N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133        N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19        T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22        MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31        MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905        MISSING (IN POLYMORPHIC EPITHELIAL
FT                                 ISOFORM).
FT   VARSPLIC   1077   1087        FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                 ISOFORM).
FT   VARSPLIC   1088   1255        MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2        T -> A (IN REF. 11).
FT   CONFLICT    134    134        P -> Q (IN REF. 9).
FT   CONFLICT    154    154        P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021        S -> T (IN REF. 3).
FT   CONFLICT   1251   1251        A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;
```

```
Query Match             9.2%;  Score 651;  DB 1;  Length 1255;
Best Local Similarity   27.8%;  Pred. No. 3.9e-20;
Matches  295;  Conservative   72;  Mismatches  460;  Indels  234;  Gaps   51;

Qy    181 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 239
          | |||          | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    240 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTI---K 279
          |    :||:|    |    ||                 | |     ||||     |:    |
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy    280 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 333
          || :   ||   | ||| |   ||| :|  |        || |  ||| :|  ||   |
Db    126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy    334 SAPTTPKEPAPTTP----------KKPAPTTPKEP------APTTPKEPTPTTP------ 371
          ||| |   |  | |           :||| :   |      || |   |  | |
Db    180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy    372 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 412
              ||| :  |        || |   |  |||        || |   |  | |
Db    240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy    413 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 467
          || |   |  | |      |:   |: || :|  ||   | ||| |   |:| :|  |
Db    300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy    468 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 515
              || |   |  | |      || |   |  | |        || |   |  |||
Db    357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy    516 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 575
          |:  :| |        |      || |   | | |      |        | |   | |
Db    417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy    576 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 627
          | |    :||:|   |  | |       || |   |  | |    |     :|| |   |
Db    473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy    628 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 675
          | |            :|||  ||      ||    :|||  | |     |:    ||
Db    529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy    676 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 721
          |   |  | |       :|| |   |  | |           :|||   ||    | ||
Db    584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy    722 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 765
          :|||  | |       :|| |   |  | |           :|||  :  | ||    |
Db    643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700

Qy    766 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 818
          | |   | :       :  :||     |  |    : ::|     |   | :| | |
Db    701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753

Qy    819 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 871
            |   |:  | |  | :|       |:||    :|      |       |:   |
Db    754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807

Qy    872 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 927
            | |    :|:    |  ||  |        |:  :    |  || |    |::  | |
Db    808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867
```

```
Qy     928 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 987
           |   : ||      |  | | |    | :|  |    :: |:  |                 |
Db     868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA------------PG 909

Qy     988 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 1046
           | | :  |   | :  | |:     |                ||   |  | : :      | :
Db     910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy    1047 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 1087
           : |:          | : |||        :||| ||
Db     956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;
```

```
  Query Match             7.9%;  Score 556.5;  DB 1;  Length 875;
  Best Local Similarity   28.7%;  Pred. No. 2e-16;
  Matches  272;  Conservative  106;  Mismatches  375;  Indels  195;  Gaps   56;

Qy     148 TQHNKVSTSPKITTAKPINP----------RP--SLPPNSDTSKETSLTVNKETTVET 193
           |:|  |    | : : |  |           ||  | ||    ||   | : |: :
Db       1 TKHEPVYKPVKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYG-SKTNYLPLAKKLSSYK 59

Qy     194 KETTTTNKQTSTDG--KEKTT---SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKG 248
             || | :|:    | | |   : |   |   |          |:   || ||: :
Db      60 PIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYPPTYKSKPTYKPKITYPPTYKAKPSY-- 117

Qy     249 PALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTT 299
           |:   ||:  | | |      | :  ||   | | |  |: |     |:   |:: |: |:
Db     118 PSSYKPKKTYPPTYKPKLTYPPTYKPKPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSY 177

Qy     300 PKEPAPTTTKEPA--PTTPKEPA--------PTTTKEPA-PTTTKSAPTTPK--EPAPTT 346
           |    ||   :|:  ||   :|:         ||   :|  |:| |: |: |   :  ||
Db     178 P----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTY 233

Qy     347 PKKPA--PTTPKEPA--PTTPKEPT--PTTPKEP---APTTKEPAPTTPKEPA--PTAPK 395
             ||:  ||   :|:  ||   :|:  ||   :|   |  | : ||   :|:  ||
Db     234 KAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYPPTYKA 293

Qy     396 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 445
           ||:  ||   :|:  ||   :|:  ||   :|| ||   :|: | | |: ||   :|
Db     294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy     446 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 501
           |:| |: |: |    ||   :|:      :  ||   ||: ||   :|:  ||   :|:
Db     354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy     502 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 561
             | |  ||   :|       | |:|  |  |: |    |:   ||:     :  ||
Db     410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy     562 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 615
           :||        |:| ||     |  ||     |  | | |      |:   ||:| | |
Db     458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy     616 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 671
           || |   | | | |: |    ||   | | |                 || : | |    |
Db     507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy     672 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 723
           ||     | | |       ||   :|:  ||   ||:   |   |:| |  |:   :
Db     552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy     724 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 772
           |  |: |    ||   :|:  ||   ||  |:|    |  ||     ::| | | |
Db     612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy     773 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 821
           | | :|:      |   |:|| ||   :|     :|  | |     |:        ||    |
Db     668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy     822 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 881
           | | | :   | | |  | | |  | :   |   | ::|        :  |    |
Db     727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy     882 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 938
           | |: :: |      |||   |    ||: |  |  :        |||  |     |
Db     778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826

Qy     939 TSTKKPKTMPRVRKPKTTPTP-----------RKMTSTMPELNPTSR 974
           :: |   : |   | ||:  |           :  ||  |   ||:|
Db     827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874
```

```
RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
```

```
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER      1      1
FT   DOMAIN       <1     109       MAM 1.
FT   DOMAIN       112    136       MAM 2 (PARTIAL).
FT   DOMAIN       161    326       MAM 3.
FT   DOMAIN       322    446       MAM 4.
FT   DOMAIN       483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                 (MUCIN-LIKE DOMAIN).
FT   DOMAIN       953    1065      VWFD 1 (PARTIAL).
FT   DOMAIN       1066   1454      VWFD 2.
FT   DOMAIN       1455   1861      VWFD 3.
FT   DOMAIN       1862   2292      VWFD 4.
FT   DOMAIN       2293   2684      VWFD 5.
FT   DOMAIN       ?      ?         EGF-LIKE.
FT   CARBOHYD     74     74        N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1023   1023      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1099   1099      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1618   1618      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1737   1737      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1832   1832      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1878   1878      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     2136   2136      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     2505   2505      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT     2374   2379      NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER      2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;
```

Query Match 7.9%; Score 555.5; DB 1; Length 2700;
Best Local Similarity 32.6%; Pred. No. 5.5e-16;
Matches 199; Conservative 64; Mismatches 240; Indels 107; Gaps 32;

```
Qy   164 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 223
             |:   | |||  |  |      |   |||  |:  | : |            | | ||||
Db   445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy   224 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 278
             |     ||      |  |    |: | ::| |   || | |  :   ::||     || :
Db   488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy   279 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 337
           || | |  ||     ||   |     |:|  |||| |   |  :||  ||   | ||
Db   540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy   338 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 397
           ::|:  | |   ||    |    : ::||    |   || |   || | | :  :|:|
Db   597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy   398 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 457
           ||||  |  || | |    :|::|: .| | |  :  |    | |   ||     || |
Db   648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy   458 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 512
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| | |
Db   702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy   513 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 562
           ::|  :| | | ||     :|:|||  | |||      || | :    : |:|  | | |
Db   758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy   563 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 622
             :||:|  :| |            || | |  || | |       ||  :| | | ||
Db   815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy   623 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 676
           ::|  :  |   | |    :|::||  |       ::||   | | | ||  ::
Db   856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy   677 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 734
           |  |:||   | |   :   ||:||      :  |||   :||    | |    |:
Db   912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966
```

```
Qy     735 PTTPKEPAPT 744
           | : | | |:
Db     967 PASCKSPRPS 976


RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73        H-A-P-P.
FT   REPEAT      148    151        H-A-P-P.
FT   DOMAIN      229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235        1.
FT   REPEAT      236    242        2.
FT   DOMAIN      205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             7.8%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 2.5e-16;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy     242 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 290
           || ||: | | |   :| |::     |  | :| |     |:|   : ||     | |
Db      24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy     291 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 341
           :    |:   |       |:   | |:   :|  | |:    |        |:  :
Db      81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy     342 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 396
           |       |:|:    |     | |:  | : : |:| ::   | | |  :| ||
Db     141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy     397 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 451
           |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db     194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy     452 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 505
             |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db     250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy     506 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 555
           | ||   :|  || : |   :|  | |  |::      |   :| |  :|  | |
Db     310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy     556 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 612
           | |  | | :  | ||  ::  |    :|  ||| |   | ||    | | |  :
Db     370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy     613 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 672
           |       | |   | |      | |     |  |  |       :| |  |  : |     |
Db     428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy     673 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 732
           | :|    |      | |  | |     | |    |   : || |     |:| |
Db     482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy     733 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 790
           |     | |    |: | ||  : | | |     |      |   |: |   ||
Db     539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy     791 PTPKALENSPKEPGVPTTKTPAATKP 816
           |:|           ||| :| :  |
Db     599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
```

```
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;


  Query Match             7.5%;  Score 533;  DB 1;  Length 1087;
  Best Local Similarity   27.5%;  Pred. No. 2.1e-15;
  Matches  216;  Conservative   81;  Mismatches  330;  Indels  158;  Gaps   38;

Qy     127 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 181
           :|   |     | :|      |  ||  | || :   |   :             : |||
Db     404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy     182 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 241
           :   :   :   |  |  :     |:|   : :  :  |:  |   :| ::  | |  :
Db     450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy     242 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 301
           ::  |  | :  :  :|   | ||    :  :|   || |   | |    : : | :|
Db     510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy     302 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 351
           || :|   | ||   || ||  | | |    :|:  || |   | ||    |   | ||
Db     565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy     352 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 396
               |   | ||    | | | |    :|:  | ||   :  :|   | ||    |   |
Db     620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy     397 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 448
           |    :|  || ||    | ||    :  ||   | ||    |   || |   | ||
Db     677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy     449 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 503
           : :|  :| |  ||    | ||    |   | |       : | |  || | :  :  :|
Db     730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy     504 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 563
           | |    | ||      | |  :| |     |||  |   : |   || :   | | :
Db     790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy     564 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 611
           |  ||:  :|   || :|   || | |:     ||:    :|        |||     :|
Db     849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy     612 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 669
             |||  |  : |: ||    ||  :|   | |::|  :|::  |     |    :
Db     905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964

Qy     670 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 729
           : ||    || |   ||    ||  |   |  |:|:     || |: |
Db     965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP------------ 1006

Qy     730 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 789
                 ||  |  |:|                      | :| ||  : |:|
Db    1007 --------KKEEMPAAPEK-----------------KDTKEEKTTESRKPEEK------ 1034

Qy     790 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 849
           || :|   ||     :| |  :||:     |  :|::     ||      | |:: |
Db    1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy     850 TTTEK 854
           |  ||
Db    1083 TKGEK 1087


RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
```

```
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;


  Query Match             7.5%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 2.9e-15;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy     258 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 316
           ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db     599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy     317 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 363
           :  |  |   |  ::  | |:|| :  | :||  |       ||:|| : :    |:||
Db     652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy     364 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 422
           : :   ||  || ::    |:|| : :       | :||  || ::     |     :
Db     712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy     423 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 478
             ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db     765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy     479 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 533
           ||    | :||  |  ::      |    | : |  || ::   | |:||       ||
Db     818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy     534 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 583
           :|| :  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db     877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy     584 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 643
           |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db     936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS------AHSTPSTPVDSSAHST 983

Qy     644 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 702
                || :|  | :|     |:|| :  | :||  || ::   |  | :  | :||
Db     984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy     703 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 758
           |     | ::  |  || :|  | :||      :| :||  || ::    |  || ||
Db    1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy     759 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 817
           :::  |||:|  : :: | :|  ||| : ||  ||                  |||  :
Db    1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy     818 MT 819
           |
Db    1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
```

```
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;

  Query Match             7.5%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity  26.2%;  Pred. No. 2.4e-15;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy     229 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 284
           :| |  :| | ||    |      ||  |        |  | : ||  | ||     | |
Db       8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy     285 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 339
             |||    :|  ||    ||     | |     || :   |                 ||| |
Db      62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy     340 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 385
           |  ||    |        | |   :  ||   ||        : |  || ||   |
Db     117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy     386 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 437
             ||  |      ||      ||   : |||    | |  | |:  ||     :   :
Db     177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy     438 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 493
           |    ||   ||  |::    |   ||  :  ||  |    |   : : |          ||
Db     237 PLAAAEPVVVAPPATETPVVAPAAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy     494 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 534
           ||         : |  |     ||   ||   ||    ||         | |||  : |
Db     297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy     535 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 590
             | : ||     | | |   || | |:      || |   ||:   :|   || : ||
Db     357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy     591 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 639
                 |    || |     |  | ||   :|| |:           ||:     ||
Db     414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy     640 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 691
             ||  :  |  ||:  |  | |    || |       | |  ||  |    ||   |
Db     470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy     692 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 737
           | |       |  |  | |  |  ||:     :   ||:|  | | |     | |   :
Db     528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy     738 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 787
             ||  |  :| |       | |: :  : |   : |  | :: |           | ||
Db     586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy     788 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 845
           : |  |:|:     :| :  |: |  |:        |      :: |||  |  ||| :::
Db     642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy     846 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 898
                 :     :|       | |        | | | :    : :  | | ::
```

```
Db     698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy     899 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 958
                 : |: |:        : |  |  : ||   : |  :  |      |   | ||
Db     758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy     959 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 1001
             | : :|   |   :   :|:   |:     ||    :| |:      |||
Db     814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864


RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL       1     20       POTENTIAL.
FT   CHAIN       21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN      21     41       NONREPETIVE LINKER.
FT   DOMAIN     124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                               P-[PST]-[ST].
FT   DOMAIN     184    192       NONAPEPTIDE 1.
FT   DOMAIN     213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;

  Query Match             7.3%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity  29.2%;  Pred. No. 7e-15;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy     125 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 175
             |   | ||   | :| | | :    ::|       ||: | | |  :|   | | |
```

```
Db      29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy     176 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 235

Db      85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy     236 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 294

Db     140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy     295 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 349

Db     186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy     350 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 402

Db     237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy     403 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 457

Db     293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy     458 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 507

Db     351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy     508 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 562

Db     411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy     563 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 613

Db     470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy     614 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 670

Db     530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy     671 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 726

Db     584 --------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy     727 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 775

Db     629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy     776 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 831

Db     688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy     832 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 891

Db     736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy     892 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 946

Db     786 I--------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy     947 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 988

Db     828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871
```

RESULT 12
MUC1_XENLA
ID MUC1_XENLA STANDARD; PRT; 662 AA.
AC Q05049;
DT 01-OCT-1994 (Rel. 30, Created)
DT 01-OCT-1994 (Rel. 30, Last sequence update)
DT 01-OCT-1994 (Rel. 30, Last annotation update)
DE INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS Xenopus laevis (African clawed frog).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC Xenopodinae; Xenopus.
OX NCBI_TaxID=8355;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Skin;
RX MEDLINE=93077556; PubMed=1447205;
RA Hauser F., Hoffmann W.;
RT "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT polymorphism.";
RL J. Biol. Chem. 267:24620-24624(1992).
CC -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC -!- TISSUE SPECIFICITY: SKIN.
CC -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC ---

CC ---
DR EMBL; L02115; AAA74725.1; -.
DR PIR; A45155; A45155.
DR HSSP; P04002; 1WFA.
DR InterPro; IPR000519; P_trefoil.
DR Pfam; PF00088; trefoil; 6.
DR SMART; SM00018; P; 6.
DR PROSITE; PS00025; P_TREFOIL; 6.
KW Repeat; Amphibian skin; Glycoprotein; Alternative splicing.

FT	NON_TER	1	1	
FT	DOMAIN	81	144	8 X 8 AA APPROXIMATE TANDEM REPEATS, ALA/THR-RICH.
FT	REPEAT	81	88	1-1.
FT	REPEAT	89	96	1-2.
FT	REPEAT	97	104	1-3.
FT	REPEAT	105	112	1-4.
FT	REPEAT	113	120	1-5.
FT	REPEAT	121	128	1-6.
FT	REPEAT	129	136	1-7.
FT	REPEAT	137	144	1-8.
FT	DOMAIN	161	202	P-TYPE 1.
FT	DOMAIN	218	301	8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT	REPEAT	218	224	2-1.
FT	REPEAT	225	239	2-2.
FT	REPEAT	240	249	2-3.
FT	REPEAT	250	259	2-4.
FT	REPEAT	260	275	2-5.
FT	REPEAT	276	287	2-6.
FT	REPEAT	288	294	2-7.
FT	REPEAT	295	301	2-8.
FT	DOMAIN	306	347	P-TYPE 2.
FT	DOMAIN	353	394	P-TYPE 3.
FT	DOMAIN	402	522	12 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT	REPEAT	402	411	3-1.
FT	REPEAT	412	419	3-2.
FT	REPEAT	420	431	3-3.
FT	REPEAT	432	443	3-4.
FT	REPEAT	444	453	3-5.
FT	REPEAT	454	460	3-6.
FT	REPEAT	461	472	3-7.
FT	REPEAT	473	479	3-8.
FT	REPEAT	480	491	3-9.
FT	REPEAT	492	498	3-10.
FT	REPEAT	499	515	3-11.

```
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;
```

Query Match 7.2%; Score 509.5; DB 1; Length 662;
Best Local Similarity 28.1%; Pred. No. 1.3e-14;
Matches 223; Conservative 48; Mismatches 276; Indels 247; Gaps 30;

```
Qy    198 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 257
Db      3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy    258 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 317
Db     52 --------AAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy    318 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 375
Db    104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy    376 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 420
Db    164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy    421 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 480
Db    219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy    481 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 540
Db    246 ------------TTTTKATPTTT------------TTTKATTTTT----TPTT----- 270

Qy    541 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 600
Db    271 ---------TTTTTKATTTPTTTTTTTP------------------------------ 289

Qy    601 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 660
Db    290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy    661 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 710
Db    329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy    711 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 770
Db    385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTPTTT 436

Qy    771 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 830
```

```
Db    437 TTTTT----------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy    831 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 890
Db    464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy    891 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 946
Db    511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy    947 MPRVRKPKTTPTPR 960
Db    567 LPQVADCKVAPSSR 580
```

```
RESULT 13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
```

```
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;


  Query Match             7.1%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 5.5e-14;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     257 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 310
           |:|     || |   :  ||     :|:      |     :|:  :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     311 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 368
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :|::|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     369 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 428
           |:|   :||:   :||:|   :||:|   :||:|   :|||:|            |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     429 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 488
           :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     489 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 548
             :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db    1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     549 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 606
           ||:|   :||:|   :|||:|   :|||:| : :|::|:     ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     607 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 665
           :|   :|||:|| |          ||:|   |:  |  ||: |   | || | :|||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy     666 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 724
           ||:|   :|||||  :||:|  |   |      ||:||          :|||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     725 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 782
           :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     783 STPE 786
           |  |
Db    1966 SDEE 1969


RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;


  Query Match             7.1%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 8.7e-14;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;

Qy     257 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 310
           |:|     || |   :  ||     :|:      |     :|:  :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     311 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 368
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :|::|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624
```

```
Qy     369 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 428
           |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     429 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 488
           :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     489 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 548
             :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db    1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     549 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 606
           ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     607 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 665
           :|   :||:|| |          ||:|   |:  |  | : |   | || | :||:|| :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy     666 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 724
           ||:|   :|||||  :||:|  |   |     ||:||        :||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     725 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 782
           :||:|   :||:||    :||:|   :||:|        |  :::|           |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP-----TYSLTSPAI---------SPDD 1965

Qy     783 STPE 786
           |  |
Db    1966 SDEE 1969
```

```
RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER      1      1
FT   DOMAIN     276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT   164    164       L -> I (IN REF. 2).
FT   CONFLICT   185    185       I -> S (IN REF. 2).
FT   CONFLICT   193    193       L -> T (IN REF. 2).
FT   CONFLICT   199    199       M -> T (IN REF. 2).
FT   CONFLICT   346    346       K -> N (IN REF. 1).
FT   CONFLICT   373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT   482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT   485    485       P -> S (IN REF. 2).
FT   CONFLICT   570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT   591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT   727    727       A -> V (IN REF. 4).
FT   CONFLICT   757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT   769    769       T -> L (IN REF. 2).
FT   CONFLICT   775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;
```

```
  Query Match             7.0%;  Score 497;  DB 1;  Length 831;
  Best Local Similarity   27.6%;  Pred. No. 5e-14;
  Matches  195;  Conservative   59;  Mismatches  307;  Indels  146;  Gaps   35;

Qy     143 PDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQ 202
           |  ||  | || : ||   :                : ||| :   :   ::  |  | :
Db     197 PSMST--HIKVKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEED 242

Qy     203 TSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----T 258
```

```
Db    243 KEAQGEEEEEAEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKS 293

Qy    259 PTTPKEPA-STTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTK 309

Db    294 PAEAKSPAEAKSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVK 352

Qy    310 EP----APTTPKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA-- 351

Db    353 SPGEAKSPAEAKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATV 411

Qy    352 --PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTA 393

Db    412 KSPVEAKSPAEVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVV 470

Qy    394 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEP 444

Db    471 AKSPAEA--KSPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSP 528

Qy    445 APTTTKS-----APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTT 495

Db    529 VKEGAKSLAEAKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPE 588

Qy    496 PKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEEL 555

Db    589 AKPPAKEEAKRPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV 644

Qy    556 APTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTT 615

Db    645 ---KAKEPPKKVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEA 700

Qy    616 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPT 675

Db    701 KKEEA----KEKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKP 744

Qy    676 TPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAP 735

Db    745 SEKE----KPK-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAK 784

Qy    736 TTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 782

Db    785 AKEED-------KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824
```

Search completed: April 26, 2002, 16:33:31
Job time: 641 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:26:24 ; Search time 61.21 Seconds
(without alignments)
481.977 Million cell updates/sec

Title: AA8
Perfect score: 7064
Sequence: 1 MAWKTLPIYLLLLLSVFVIQ.....ARAITTRSGQTLSKVWYNCP 1311

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

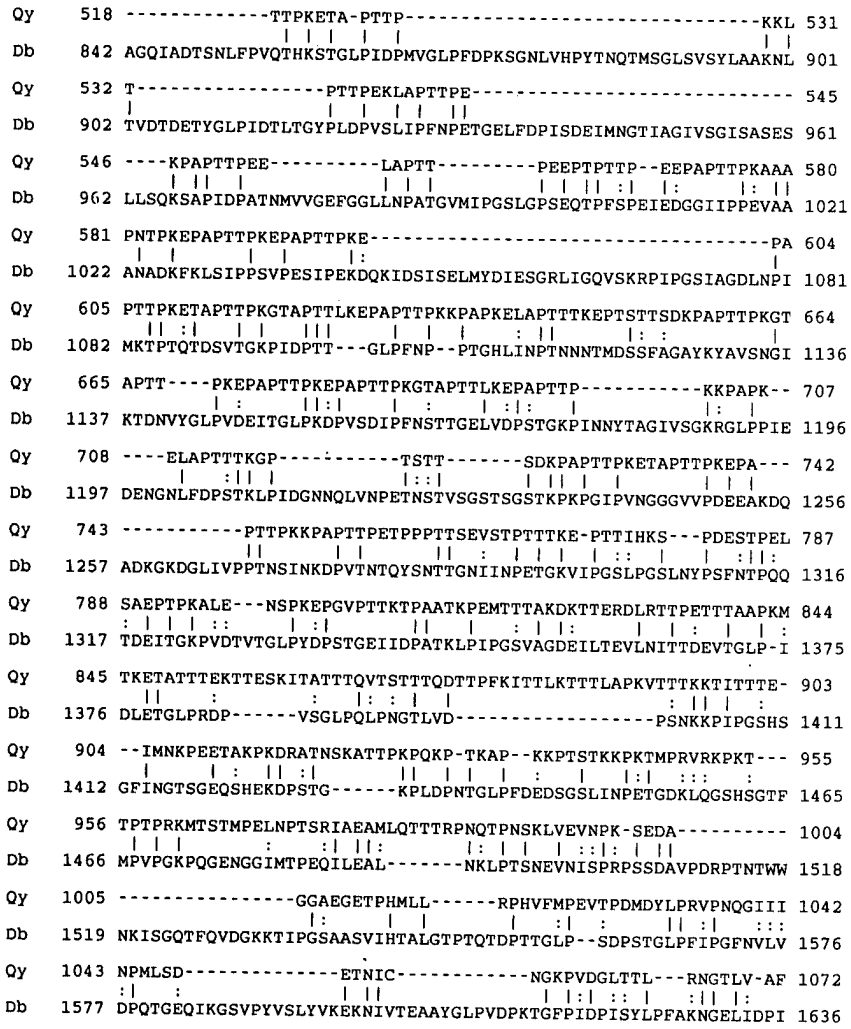
SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	7.4	805	4	US-09-103-429A-4	Sequence 4, Appli
2	508	7.2	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	7.2	744	6	5202236-25	Patent No. 5202236
4	498.5	7.1	786	4	US-09-103-429A-3	Sequence 3, Appli
5	489	6.9	1721	3	US-08-700-651-5	Sequence 5, Appli
6	489	6.9	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	6.9	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	6.9	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	6.7	652	6	5202236-13	Patent No. 5202236
10	452	6.4	960	4	US-09-219-849-5	Sequence 5, Appli
11	424.5	6.0	1867	2	US-08-479-537A-5	Sequence 5, Appli
12	424.5	6.0	1867	4	US-09-083-116-5	Sequence 5, Appli
13	424.5	6.0	2035	2	US-08-479-537A-2	Sequence 2, Appli
14	424.5	6.0	2035	4	US-09-083-116-2	Sequence 2, Appli
15	419.5	5.9	2476	2	US-08-276-967-2	Sequence 2, Appli
16	419	5.9	1185	4	US-09-041-886-23	Sequence 23, Appl
17	417	5.9	829	1	US-08-642-255-132	Sequence 132, App
18	417	5.9	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	5.9	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	5.9	837	1	US-08-477-509B-103	Sequence 103, App
21	417	5.9	837	1	US-08-642-255-101	Sequence 101, App
22	417	5.9	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	5.9	837	3	US-08-482-085B-103	Sequence 103, App
24	417	5.9	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	5.9	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	5.9	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	5.7	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	5.7	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	5.7	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	5.6	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	5.6	682	1	US-08-397-633A-36	Sequence 36, Appl
32	381	5.4	1848	4	US-08-296-791-6	Sequence 6, Appli
33	381	5.4	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
34	376	5.3	960	4	US-09-219-849-6	Sequence 6, Appli
35	375.5	5.3	1231	3	US-08-904-263A-4	Sequence 4, Appli
36	374.5	5.3	1537	1	US-08-325-267A-2	Sequence 2, Appli
37	371	5.3	761	2	US-08-707-237A-84	Sequence 84, Appl
38	371	5.3	762	1	US-08-642-255-114	Sequence 114, App
39	371	5.3	762	1	US-08-397-633A-26	Sequence 26, Appl
40	370.5	5.2	1064	1	US-08-642-255-62	Sequence 62, Appl
41	368	5.2	762	1	US-08-642-255-120	Sequence 120, App
42	368	5.2	762	1	US-08-397-633A-31	Sequence 31, Appl
43	367.5	5.2	1187	1	US-08-320-559-28	Sequence 28, Appl
44	367.5	5.2	1187	3	US-08-545-860D-28	Sequence 28, Appl
45	367.5	5.2	1187	5	PCT-US94-04496-28	Sequence 28, Appl

ALIGNMENTS

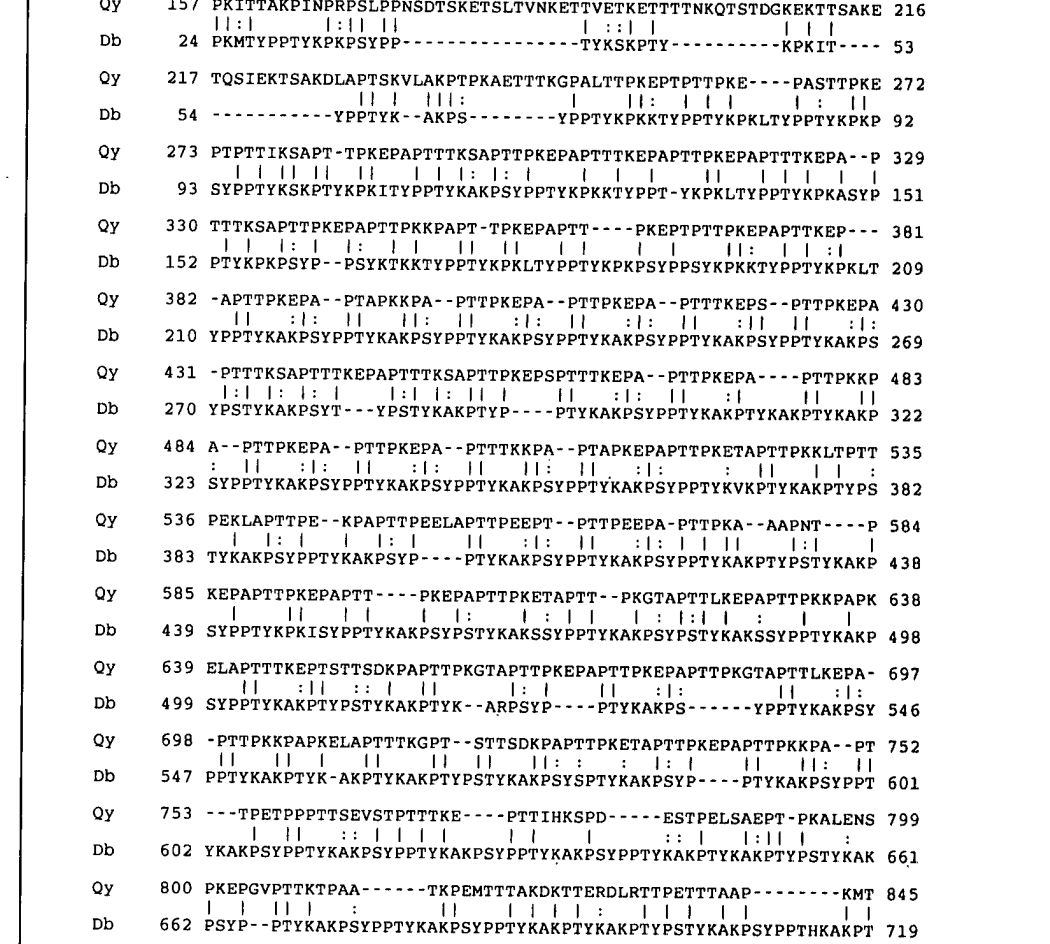
```
RESULT   1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-4


  Query Match             7.4%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity   27.8%;  Pred. No. 1.5e-25;
```

```
  Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy    231 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 285
          |:  |    |:     | |||  |    |  | |     | |   |                |
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    286 KEPAPTT------------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 333
          :: || |              :  | |  || ||    || | :  |||||:  |||||
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    334 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 393
           ||||  :   ||           |||| : || |     | ||: |  ||
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    394 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 453
              ||||  : ||||  : |||||:          |||||:.|||||:  ||||   |
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    454 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 497
          |||   :|  |  || |||    || |||  ||||:    ||  |
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    498 ------------------------------------------------EPAPTTTK 505
                                                           |||  |:
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    506 K----------------PA---------------------------------------- 508
                             ||
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    509 --------------------------------PTAP--KEPAPT-----------TPKET 523
                                           || |   || |            | |:
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    524 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 549
                                  :| |    :| |            :   |     ||
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    550 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 609
          |     |||       |:|   || | |  |||  |    ||    ||||     ||||
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    610 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 666
            :|||   |  |||     ||||    |  |:  | |  ||:  :   |||      ||
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    667 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 725
          ||     |||      |:|   ||||   ||| |           | |  ||:  |:  |
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    726 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 785
          ||     |  | |   ||||   ||| |  | ||| :  : |     ||  |       |
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    786 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 817
            :| |              | ||  || | ||
Db    686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             7.2%;  Score 508;  DB 3;  Length 1837;
  Best Local Similarity   21.3%;  Pred. No. 3.1e-24;
  Matches  295;  Conservative  100;  Mismatches  494;  Indels  496;  Gaps   48;

Qy    139 KVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTT 198
          | ||  |:||     :|:   ||        :    : |:  |: |    ||  |  |||
Db    307 KHTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 366

Qy    199 TNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPT 258
          |   |:|  |: ||:   | :   |:       |:       |    |||     || |  |
Db    367 TTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 426

Qy    259 PTTPKEPASTTPKEPTPTTIKSAPTTPKEP--APTTTKSAPTTPKEPAPTTTKEPAPTTP 316
            ||     :||    | ||  :  || |:|    ||| :  ||      |||     ||
Db    427 TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 486

Qy    317 KEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE---------- 365
            | ||||:|  ||| :  || |:|  ||      || :  :   | |
Db    487 TTTATTTTKKPTTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECE 546

Qy    366 ------------------------PTPTT---------------------------PK 372
                                      |   |                         |
Db    547 AKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPD 606

Qy    373 EPAPTTKEP--------------------APTTPKEPA-PTAPKKP------------ 397
              |  |                      :|  |: :  :|| :|
Db    607 RSLDFTIPPVAGHNSCSIIVGVSGGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCT 666

Qy    398 ----------------APTTPKEPAPTT---PKEPAPTTTKEPS-PTTPKEPAPTTTKS- 436
                           ||  :|  ||   | :|  |||  || |||      ||| :
Db    667 AKYGAIHSGYQTSADFVTTTTAKPTTTTTGAPGQPTTTTTGSPSKPTTTTTTKATTTTTI 726

Qy    437 ----APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPA 492
                 |||::|  |||    |  |   : ||| :|  ||     | ||     |||
Db    727 LNPIITTTTQKPTTTTTTKVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT----- 781

Qy    493 PTTPKEPAPTTTKKPAP-------------------------------TAPKEPAP--- 517
            || |    |||  | |                                 :  || |
Db    782 TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQ 841
```



RESULT 3
5202236-25
;Patent No. 5202236
; APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
; TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
; NUMBER OF SEQUENCES: 39
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/528,762
; FILING DATE: 25-MAY-1990
; APPLICATION NUMBER: 82,456
; FILING DATE: 07-AUG-1987
; APPLICATION NUMBER: 933,945
; FILING DATE: 24-NOV-1986
; APPLICATION NUMBER: 650,128
; FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
; LENGTH: 744
5202236-25

Query Match 7.2%; Score 506.5; DB 6; Length 744;
Best Local Similarity 29.9%; Pred. No. 1.3e-24;
Matches 239; Conservative 78; Mismatches 308; Indels 175; Gaps 50;



RESULT 4
US-09-103-429A-3

```
; Sequence 3, Application US/09103429A
; Patent No. 6187558
; GENERAL INFORMATION:
;   APPLICANT:  Granados, Robert R
;   APPLICANT:  Wang, Ping
;   TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;   TITLE OF INVENTION:  cDNA and Related Products and Methods
;   NUMBER OF SEQUENCES:  4
;   CORRESPONDENCE ADDRESS:
;     ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;     STREET:  118 No. 6187558th Tioga
;     CITY:  Ithaca
;     STATE:  NY
;     COUNTRY:  USA
;     ZIP:  14850
;   COMPUTER READABLE FORM:
;     MEDIUM TYPE:  Floppy disk
;     COMPUTER:  IBM PC compatible
;     OPERATING SYSTEM:  PC-DOS/MS-DOS
;     SOFTWARE:  PatentIn Release #1.0, Version #1.30
;   CURRENT APPLICATION DATA:
;     APPLICATION NUMBER:  US/09/103,429A
;     FILING DATE:  24-JUN-1998
;     CLASSIFICATION:  800
;   ATTORNEY/AGENT INFORMATION:
;     NAME:  Michaels, Christopher A
;     REGISTRATION NUMBER:  34,390
;     REFERENCE/DOCKET NUMBER:  BTI-39
;   TELECOMMUNICATION INFORMATION:
;     TELEPHONE:  (607) 256-2000
;     TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;   SEQUENCE CHARACTERISTICS:
;     LENGTH:  786 amino acids
;     TYPE:  amino acid
;     STRANDEDNESS:  single
;     TOPOLOGY:  linear
;   MOLECULE TYPE:  protein
;   HYPOTHETICAL:  NO
;   ANTI-SENSE:  NO
;   ORIGINAL SOURCE:
;     ORGANISM:  Trichoplusia ni
;     TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             7.1%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 4.6e-24;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    238 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 297
          | | || ||: || |     ||| : | :||          ||||  :   |||  ||
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    298 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 357
          || :  |||||:  | ||    ||||| : ||||| ||||  : |||| :  |||| :
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    358 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 414
           |||| : ||  |     |    || |||    | |   || |||  ||||: |  |
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    415 ----------------------------------------------PTTTKEPSPTTPK- 427
                                                         |   :  ||   :
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    428 ----EPAPTTT---------------------------------------------------- 434
               |||  |
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    435 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 467
                                     :|      | |  |    || | |  |     |
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    468 APTT-----PKEP----------------------APTTPKKPAPTT---------- 487
          |  :     | |                        :| |    :| |
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    488 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 544
               |||     |||     ||||   |:   |  | | |   : |||     ||
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    545 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 602
            ||||     ||||     :|||   | ||    ||| |  |: |   |    |||
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    603 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 662
           |||     ||||     |||:  |    |   ||   || ||             | |
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT------------VTVPP 632

Qy    663 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 722
            ||||   ||| |     ||     || |:                || ||  | :|||
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    723 DKPAPTTPKETAPTTPKEPAPTTPKKP 749
           || |  |     |  | |  |  |
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   5
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5


  Query Match             6.9%;  Score 489;  DB 3;  Length 1721;
  Best Local Similarity   22.1%;  Pred. No. 4.5e-23;
  Matches  284;  Conservative   82;  Mismatches  476;  Indels  442;  Gaps   41;

Qy    111 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 170
          : | ||   | :|     ||    | : | | |  :|   | :   |||          :
Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy    171 LPPNSDTSKE---------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 215
           | ||: | |               |     | ||  |  ||||   |:|
Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTT---------- 208

Qy    216 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 275
                                  |||     ||    | ||     :||    |
Db    209 ---------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    276 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 335
          ||  :  ||      ||| :  ||      |||     ||     |||     ||| :
Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    336 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 395
```

```
Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT---------- 350

Qy    396 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 455
Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTTKKPTTTTTATTTTT 406

Qy    456 PKE-----------------------------------------------PSPTT------- 463
Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    464 ----------------------------TKEP--------------------APTTPKE 474
Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    475 PAPTTPKKPAPTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 516
Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    517 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 576
Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    577 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 636
Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTPL 680

Qy    637 PK----ELAPTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 676
Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    677 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 733
Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    734 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 779
Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    780 -------------------------PDEST-------------PELSAEPTPKALENSPK 801
Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    802 EPGVPTTKTPAATKPEMTTTAKDKTTER-----------------DLRTTPETTTAAPKM 844
Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    845 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 903
Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    904 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 948
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    949 ---------------RVRKPKTTPTPRKMTS--TMPE---------------------- 968
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    969 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 1023
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy   1024 EVTPD--------MDYLPRVPNQGIIINP--------------MLSDETNICNGK----- 1056
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy   1057 ----------PVDGLTTLRNGTLV 1070
Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281

RESULT   6
```

```
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6


  Query Match             6.9%;  Score 489;  DB 3;  Length 1721;
  Best Local Similarity   22.1%;  Pred. No. 4.5e-23;
  Matches  284;  Conservative   82;  Mismatches  476;  Indels  442;  Gaps   41;

Qy    111 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 170
Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy    171 LPPNSDTSKE--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 215
Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTT---------- 208

Qy    216 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 275
Db    209 ----------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    276 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 335
Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    336 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 395
Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT---------- 350

Qy    396 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 455
Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTTKKPTTTTTATTTTT 406
```

```
Qy    456 PKE-----------------------------------------------PSPTT------- 463
Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    464 ----------------------------TKEP--------------------APTTPKE 474
Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    475 PAPTTPKKPAPTTPKE------------------PAPTTPKEPAPTTTKKPAPTAPKEPA 516
Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    517 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 576
Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    577 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 636
Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTPL 680

Qy    637 PK----ELAPTTTKEPTS---------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 676
Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    677 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 733
Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    734 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 779
Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    780 ------------------------PDEST-------------PELSAEPTPKALENSPK 801
Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    802 EPGVPTTKTPAATKPEMTTTAKDKTTER-----------------DLRTTPETTTAAPKM 844
Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    845 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 903
Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    904 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 948
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    949 ---------------RVRKPKTTPTPRKMTS--TMPE---------------------- 968
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    969 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 1023
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy   1024 EVTPD--------MDYLPRVPNQGIIINP-------------MLSDETNICNGK----- 1056
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy   1057 ----------PVDGLTTLRNGTLV 1070
Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281
```

```
RESULT   7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/07/638,431
;      FILING DATE:  19910110
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avrom D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-07-638-431-2
```

Query Match 6.9%; Score 488.5; DB 1; Length 826;
Best Local Similarity 26.4%; Pred. No. 2.1e-23;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    274 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 326
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    327 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 382
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    383 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 440
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    441 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 498
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    499 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 557
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    558 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 616
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    617 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 676
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    677 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 736
Db    607 PKEPI-------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649
```

```
Qy    737 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 790
          |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    791 PTPKALENSPK-------------------EPGVPTTK 809
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene encoding
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  PCT/US92/00018
;      FILING DATE:  19920103
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avram D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
PCT-US92-00018-2


  Query Match             6.9%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 2.1e-23;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    274 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 326
          ||  ::  |  |  |  |       |: |:||  |    :|     |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNN 332

Qy    327 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 382
          |      :  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | :|:     |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    383 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 440
             ||   |  ||  || |  |    |: | :| |       |: | :| |    : |:
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    441 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 498
            :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    499 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 557
```

```
          |:       |  |  |  |: | |  |: |||                  |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    558 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 616
          : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    617 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 676
                    |: |::| |:|  |:  |||    |:   |  |:   |  |  |  :
Db    565 ------------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    677 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 736
          ||||               |  : ||:|    : |   : |     :   |       |
Db    607 PKEPI-------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    737 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 790
          |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    791 PTPKALENSPK-------------------EPGVPTTK 809
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   9
5202236-13
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;      LENGTH: 652
5202236-13


  Query Match             6.7%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 9.2e-23;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy    157 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 216
          ||:|      |:|| ||                | ::| |           | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    217 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 272
                    || |  |||:          |    ||:  | | |     | :  ||
Db     41 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy    273 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 329
            | | || ||  ||   | |.|: |: |    |  |  |    ||   | | |  |  |
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    330 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 384
            | |  |: |  |:  | |   ||  ||   | |    |  |    | :  | | :|  |
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196

Qy    385 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 435
            |    ||   ||:  ||   :|:  ||   :|:   | :  |: | |     |   | | |
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    436 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 482
          : ||   :|        |:| |: || |    ||   :|:  ||   :|     ||   |
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308
```

```
Qy    483 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 538
          |:  ||   :|:  ||   :|:   | |  |: |    ||   : :     |: ||   |
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    539 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 593
            || |   |  |: |    ||   :|:  ||   :|: | | | | |: |    ||
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    594 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 646
          :|  |:|    |  |    || : | |    | : |:| | :   |   |     ||
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    647 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 704
          :||  :: |  ||       |: |    ||   :|:         ||   :|:  ||   ||
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    705 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 757
            |  ||    |:   : |  |: |    ||   :|:  ||   ||  |:|    | |
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    758 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 817
          |   :: | | |       :|:      |   |:|| ||      :|  |:|      ||
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    818 MTTTAKDKTTERDLRTTPETTTAAP 842
            | | | |  :   : | |  | |
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645


RESULT  10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
;  APPLICANT:  VAN HEERDE, GEORGE V.
;  APPLICANT:  VAN RIJN, ALEXIS C.
;  APPLICANT:  BOUWSTRA, JAN B.
;  APPLICANT:  DE WOLF, FREDERIK A.
;  APPLICANT:  MOOBROEK, ANDREAS
;  APPLICANT:  WERTEN, MARC W.T.
;  APPLICANT:  WIND, RICHELE D.
;  APPLICANT:  VAN DEN BOSCH, TANJA J.
;  TITLE OF INVENTION:  SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
;  TITLE OF INVENTION:  SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
;  TITLE OF INVENTION:  PREPARATION THEREOF
;  FILE REFERENCE:  2728-2
;  CURRENT APPLICATION NUMBER: US/09/219,849
;  CURRENT FILING DATE:  1998-12-23
;  NUMBER OF SEQ ID NOS:  50
;  SOFTWARE:  PatentIn Ver. 2.1
; SEQ ID NO 5
;   LENGTH: 960
;   TYPE: PRT
;   ORGANISM: Artificial Sequence
;   FEATURE:
;   OTHER INFORMATION: Description of Artificial Sequence: Illustrative
;   OTHER INFORMATION: amino acid sequence
US-09-219-849-5


  Query Match             6.4%;  Score 452;  DB 4;  Length 960;
  Best Local Similarity   25.6%;  Pred. No. 5.1e-21;
  Matches  173;  Conservative   60;  Mismatches  309;  Indels  134;  Gaps   41;

Qy    238 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 295
          | |     :: |    |  | | |  | | |  |      : |   ::| |      :
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    296 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 351
            |  ::| |     || |  ::| |     ||        |  | | |  |   : |
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    352 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA---------------- 390
          |  |  | |  |   ::| |      ||  | :::| |     ||
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    391 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 436
                  |  ||  | |  |   ::| |     || |  :::| |  |  | |    :
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    437 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 493
            |  :::| |       | |    |  :::| |  |  | | |   :  | |  |
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    494 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 543
            |   ::| |     ||       | |  | | |  |  :    |   | |   |
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    544 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 598
            | |   |    |    :| |  | | | |  | :  | |  | | |  |   ::| |
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    599 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 641
            ||        |  ||    |  |||           ||     |   | | ||
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    642 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 696
          |  :::|    :  | |  | |:  |  | |  |  |   ::| |    | | |   ::|
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    697 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 748
            |  | | |       ::  |    ||     : |   | |     |   ::| | |
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    749 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 799
          | |  |    :  ||     : |  :::|     :|  : |  | :| |
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    800 PKEPGVPTTKTPAATK 815
            ::|| |    ||  |
Db    696 SRDPGPPGAHGPAGPK 711


RESULT  11
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
```

```
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             6.0%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 6e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    181 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 239
          | |||           | :  :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    240 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 282
          |    :||:|    |    ||               | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    283 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 323
              :||| :|        |||      |             ||      |  | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    324 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 370
              :  | | :|        |||      |  | |       ||      |  | |     |:

Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    371 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 417
            :  | | :  |        ||      |  |||        ||      |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    418 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 466
            :  | |   ||       |||      | |          |||      |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    467 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 518
           ||      | | |       ||      | | |        ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    519 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 574
          :  :  |    | |      |:    ||      |  | |      |:  :  |     |  |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    575 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 629
           | |     :||:    |  | |      |:  :  |  |  |||     | ||      |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    630 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 677
           | |           :| |   ||       ||     :| |  | |      |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    678 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 723
            | | |      :||      |  | |            :| |   ||    | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    724 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 767
           :| |  | |       :||      |  | |            :| | :  | |      | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    768 TTTKEPTT--------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 813
              | :        :  :||      |  |    : ::|     | : | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    814 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 869
           :|   :||       |:  | |  | :|  ||      |:    :             |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    870 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 929
                |    |  ||       :  :    |  ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    930 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 989
           : ||     |  |    |      | |    :|| |:  |                  | |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    990 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1048
           | :  |   |      |:     |           ||   |  |   :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1049 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1103
           |:  : :|  | |    :|   |          |   | |         |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  12
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
```

```
;       STREET:  P.O. Box 1404
;       CITY:  Alexandria
;       STATE:  Virginia
;       COUNTRY:  United States
;       ZIP:  22313-1404
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.0, Version #1.30
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  US/09/083,116
;       FILING DATE:
;       CLASSIFICATION:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  08/479,537
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5
```

```
  Query Match             6.0%;  Score 424.5;  DB 4;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 6e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    181 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 239
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    240 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTPKEPTPTTIKSAP 282
          |    :||:|    |    ||              | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    283 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 323
             :||| :|         |||      |            ||     | | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    324 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 370
            :  | | :|        |||      |  | |       ||      |  | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    371 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 417
            :  | | :  |        ||      |  |||       ||      |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    418 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 466
            :  |  |   ||      |||      |  |        |||     |  |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    467 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 518
           ||      |  | |       ||      |  | |       ||      |  |||     |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    519 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 574
          :  :  |    | |     |:    ||      |  | |     |:  :  |     |  |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    575 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 629
           | |     :||:    |  | |     |:  :  |  |  |||     | ||     |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    630 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 677
           | |          :| |   ||      ||    :| |  | |     |:    ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    678 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 723
           |  | |      :||     |  | |           :| |   ||    | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    724 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 767
           :| |  | |      :||     |  | |           :| | :  | |      | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    768 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 813
               | :       :  :||      |  |    : ::|     | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    814 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 869
           :|   :||       |:  | |  | :|  ||     |:    :            |||
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    870 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 929
                |       |  |  ||      :  :    |  ||      |::    | |
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    930 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 989
            : ||      |  |   |       | |    :||  |:  |              |  |
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    990 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1048
           | :  |   |       |:      |           ||   |  |   :      |
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1049 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1103
```

```
           |:  : :|  | |    :|   |          |   | |        |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT 13
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2


  Query Match             6.0%;  Score 424.5;  DB 2;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 6.6e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    181 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 239
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    240 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 282
          |    :||:|    |    ||               | |     ||||     |   |||
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    283 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 323
             :||| :|          |||      |             ||      |  | |     |:
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    324 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 370
            :  | | :|         |||      |  | |        ||      |  | |     |:
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    371 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 417
            :  | | :  |         ||      |  |||        ||      |  | |     |:
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    418 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 466
            :  | |   ||       |||      |  |          |||      | |
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    467 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 518
          ||      | | |       ||      |  | |        ||      |  |||      |
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    519 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 574
            :  :  |    | |      |:    ||      |  | |      |:  :  |     |  |
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    575 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 629
          | |     :||:     |  | |      |:  :  |  |  |||      | ||     |
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    630 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 677
          | |           :| |   ||      ||     :| |  | |      |:     ||
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    678 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 723
            |  | |       :||      |  | |           :| |   ||     | ||
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    724 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 767
          :| |  | |       :||      |  | |           :| | :  | |      | |
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    768 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 813
            | :           :  :||      |  |     : ::|     | :  | |
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762
```

```
Qy    814 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 869
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    870 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 929
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    930 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 989
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    990 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1048
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1049 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1103
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

```
RESULT  14
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repea
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CC
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2
```

```
  Query Match             6.0%;  Score 424.5;  DB 4;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 6.6e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy    181 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 239
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    240 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 282
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    283 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 323
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    324 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 370
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    371 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 417
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    418 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 466
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    467 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 518
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    519 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 574
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    575 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 629
```

```
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    630 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 677
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    678 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 723
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    724 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 767
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    768 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 813
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    814 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 869
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    870 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 929
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    930 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 989
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    990 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 1048
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy   1049 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 1103
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

```
RESULT 15
US-08-276-967-2
; Sequence 2, Application US/08276967
; Patent No. 5851817
; GENERAL INFORMATION:
;   APPLICANT:  Hardy, Daniel M.
;   APPLICANT:  Garbers, David L.
;   TITLE OF INVENTION:  Species-Specific Egg-Binding Proteins of
;   TITLE OF INVENTION:  Sperm
;   NUMBER OF SEQUENCES:  8
;   CORRESPONDENCE ADDRESS:
;     ADDRESSEE:  Arnold, White & Durkee
;     STREET:  P. O. Box 4433
;     CITY:  Houston
;     STATE:  Texas
;     COUNTRY:  USA
;     ZIP:  77210-4433
;   COMPUTER READABLE FORM:
;     MEDIUM TYPE:  Floppy disk
;     COMPUTER:  IBM PC compatible
;     OPERATING SYSTEM:  PC-DOS/MS-DOS
;     SOFTWARE:  PatentIn Release #1.0, Version #1.25
;   CURRENT APPLICATION DATA:
;     APPLICATION NUMBER:  US/08/276,967
;     FILING DATE:  Submitted Herewith
;     CLASSIFICATION:  530
;   ATTORNEY/AGENT INFORMATION:
;     NAME:  Kitchell, Barbara S.
;     REGISTRATION NUMBER:  33,928
;     REFERENCE/DOCKET NUMBER:  UTSD:418\KIT
;   TELECOMMUNICATION INFORMATION:
;     TELEPHONE:  713-787-1400
;     TELEFAX:  713-789-2679
;     TELEX:  79-0924
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2476 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-276-967-2


  Query Match             5.9%;  Score 419.5;  DB 2;  Length 2476;
  Best Local Similarity   32.5%;  Pred. No. 1.7e-18;
  Matches  163;  Conservative   43;  Mismatches  177;  Indels  119;  Gaps   31;

Qy    512 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 571
Db    313 PSETSVSTEKPVAPTE----KPTVPSEIYTIPTEKPMVHMEKPIVHT--EKPTVPT-EKP 365

Qy    572 APTTPKAAAPNTPKEPA----PTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 627
Db    366 TIPTEKSTVPT--KKPTVFKEPTLPPE-GPTVPAE-RPTTPPEGPAVPPKG--PTVLTE- 418

Qy    628 APTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPT-TPKEPAPTTPKEPAPTTPK 686
Db    419 ------------WPTSHTEKSTVHTEKPILPTGKSTIPTEKPMVPTKRT------TTP- 458

Qy    687 GTAPTTLKEPAPTTPKKPAPKELAPT-TTKGPTSTT---SDKPAPTTPKETAPTTPKEPA 742
Db    459 -TERTTIPAEKPTV---PIEKPMVPTERTTIPTERTTIPTEKPTVPTEKLTVPT--EKPI 512

Qy    743 PTTPKKPAPTTPETPPP-----TTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPT 792
Db    513 VPTEKPIVPTEKHTIPTEKLTVLTERTTTPTERTTIPTEKPTVPTEKPSVPT-EKPTVPT 571

Qy    793 PKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTT 852
Db    572 --------EEPTIPTEKLTVPT--ERTTTPTKRTTTPTIRTTTPTI--------RTTTPT 613

Qy    853 EKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETA 912
Db    614 ERTT------TPTIRTTTPTERTT---IPTKKTT------VPTEKTIIPT------ERTI 652

Qy    913 KPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT 972
Db    653 AP---------TTPQP-SPTLVPTQPAAVVMPST------SATTVTPRTTIASCP---PN 693

Qy    973 SRIAEAMLQTTTRPNQTPNSKL 994
Db    694 AHFERCACPVSCQ-SPTPNCEL 714


Search completed: April 26, 2002, 16:26:41
Job time: 541 sec
```

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:18:41 ; Search time 82.85 Seconds
(without alignments)
890.006 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1167
Perfect score: 5155
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....GKPVDGLTTLRNGTLVAFRG 968

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1062.5	20.6	3020	2	A43932	mucin 2 precursor,
2	942	18.3	1664	2	T18262	S-layer protein -
3	853	16.5	1489	2	T31108	cyst germination s
4	802	15.6	1274	2	T16251	hypothetical prote
5	784	15.2	1188	2	S49915	extensin-like prot
6	776	15.1	2187	2	T30826	nascent polypeptid
7	763	14.8	1367	1	S48478	glucan 1,4-alpha-g
8	671	13.0	1151	2	T18535	high molecular mas
9	667	12.9	1344	1	A35175	mucin 1 precursor,
10	659.5	12.8	1229	2	T25697	hypothetical prote
11	633	12.3	489	2	T11622	extensin class 1 p
12	632.5	12.3	7962	2	I38346	elastic titin - hu
13	632	12.3	761	2	C84672	hypothetical prote
14	629	12.2	3507	2	T34513	hypothetical prote
15	626.5	12.2	990	2	I51618	nucleolar phosphop
16	625	12.1	6642	2	T29757	protein UNC-89 - C
17	622.5	12.1	839	2	F75518	hypothetical prote
18	620.5	12.0	971	2	T19431	hypothetical prote
19	617	12.0	3570	2	T45025	mucin MUC5B, trach
20	607.5	11.8	801	2	T29018	hypothetical prote
21	607.5	11.8	924	2	S27923	gene LF3 protein -
22	605	11.7	379	2	S50125	larval glue protei
23	559.5	10.9	350	2	S22456	hydroxyproline-ric
24	556.5	10.8	856	2	T16543	hypothetical prote
25	556.5	10.8	875	2	S23760	polyphenolic adhes
26	555.5	10.8	1630	2	A53577	ascites sialoglyco
27	551	10.7	620	2	S06733	hydroxyproline-ric
28	549	10.6	2232	2	T34434	hypothetical prote
29	545	10.6	873	2	A47283	calphotin - fruit
30	542	10.5	369	2	S20500	hydroxyproline-ric
31	538.5	10.4	416	2	JU0465	extensin precursor
32	533	10.3	1087	1	QFMSH	neurofilament trip
33	532	10.3	756	2	T27642	hypothetical prote
34	530.5	10.3	1162	2	JH0557	exo-alpha-sialidas
35	530	10.3	865	2	A47282	calcium-binding pr
36	522	10.1	328	2	JQ0985	hydroxyproline-ric
37	518	10.0	813	2	S70795	vsaA protein precu
38	518	10.0	866	2	T45462	membrane glycoprot
39	518	10.0	1072	1	A37221	neurofilament trip
40	517.5	10.0	1459	2	T32271	hypothetical prote
41	512.5	9.9	867	2	T45463	membrane glycoprot
42	509.5	9.9	662	2	A45155	mucin FIM-C.1 - Af
43	505	9.8	700	2	A54641	interspersed repea
44	504.5	9.8	606	2	A43427	neurofilament trip
45	504	9.8	1832	2	T31113	mucin-like glycopr

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 20.6%; Score 1062.5; DB 2; Length 3020;
Best Local Similarity 29.2%; Pred. No. 4.2e-38;
Matches 342; Conservative 74; Mismatches 442; Indels 315; Gaps 39;

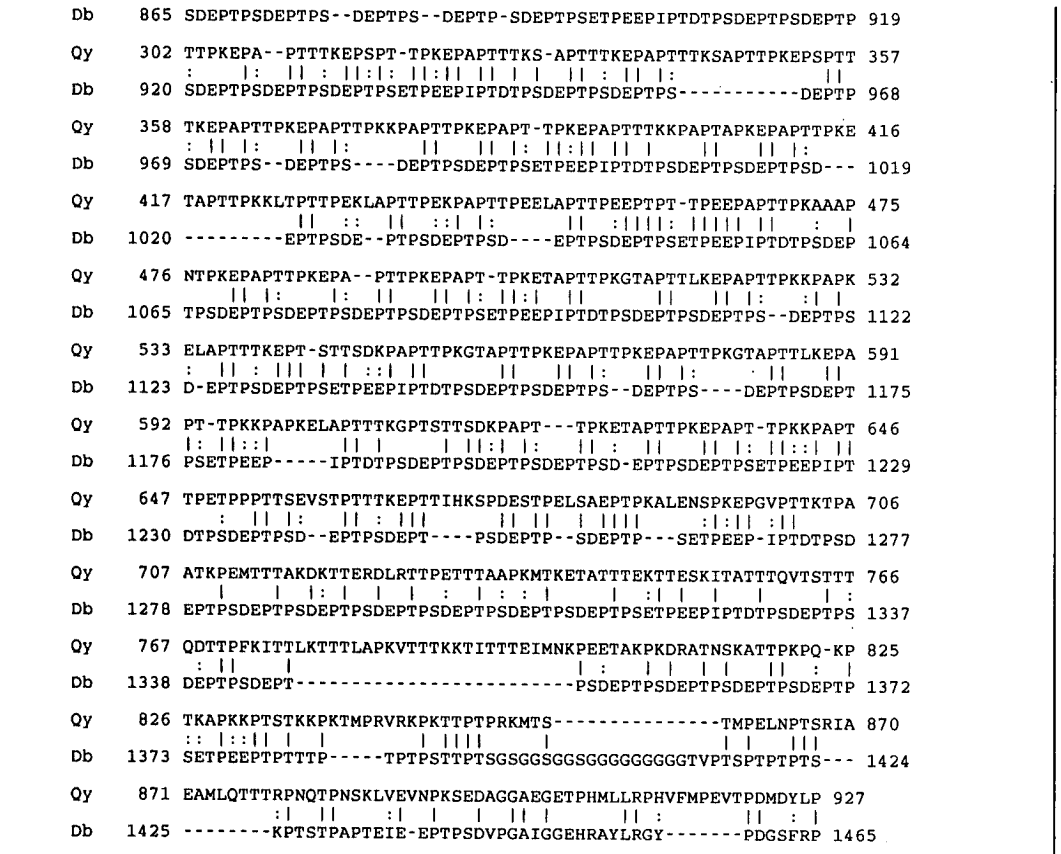
```
Qy      32 FKVTTPDTSTTQH---NKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
           :::  |: :  :|     ::| |   |       |: |        || |    ||  |
Db    1240 WEICGPNGTVEKHFNICSITTRPSTLTTFTTITLPTTP--------TSFTTTTTTTTPTS 1291

Qy      89 ET--TTTNK--------------QTSTDGKEKTTS--AKETQSIEKTSAKD-------- 121
            |  :|| |                 | ||  :         : ||  | ||
Db    1292 STVLSTTPKLCCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQH 1351

Qy     122 ----------------------------------------------LAPTSKVLAKPTPKAE 137
                                                          :   |     |:|
Db    1352 GQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPT 1411

Qy     138 TTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-TTPKEPAPTTTKSAPTTPKE 196
           |||  |  |||  || ||   | :|||  |  ||    | |||  |  |||   |||
Db    1412 TTTTLPPTTTPSPPTTTTTTPPPTTTPSPPITTTTTPLPTTTPSPPISTTTTPPPTTTPS 1471

Qy     197 PAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP-TTPKEPAPTTPKKPAPTTPKEPAPT 255
           | ||||  |  |||  |  |||  | |||| | | |||  |  :|   |  |||  |  |
Db    1472 P-PTTTPSPPTTTPSPPT-TTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTT 1529

Qy     256 TPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEP 315
           |   |  |||  |  |   | ||:     ||    | |||   | |||   |  |||  |
Db    1530 TTTPPPTTTPSPPTTTPITP-PTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSP 1588

Qy     316 SPTTPKEPAPTTTKSAP-TTTKEPAPTTTKSAP-TTPKEPSPTTTKEPAPTTPKEPAPTT 373
              |  | |||| | | |||  | |||| | | |||  |  :||  |  |||  |  ||
Db    1589 PTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTT 1648

Qy     374 PKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKL 433
              |  |||  |  |||  |  |||  |  | |  |  |||  : |||    | |||
Db    1649 TTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTP--SPPTT----TMTTPS-- 1700

Qy     434 APTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT 493
             |||  |  ||    : |||  | |||   |:|||  :    |     |||   |  ||
Db    1701 PTTTPSSPITTTTTPSSTTTP-SPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTT 1759

Qy     494 PKEPAPTTP-----KETAPTTP-----------KG---------------------- 512
           |  |: | |       | ||||           |
Db    1760 PLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAA 1819
```

```
Qy     513 ---------------------------------------------------------- 512

Db    1820 NISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECV 1879

Qy     513 TAPTTLKEPAPTTPKKPAPKELAPTT--TKEPTSTTSDKP----APT-TPKGTAPTTPKE 565
           | |||:     || : | |  :  ||  |  || |::  |    ||| ||  |  |
Db    1880 TQPTTM---TTTTTENPTPTPITTTTTVTPTPTPTSTQSPNGLQAPTPTPISTTTTVTPT 1936

Qy     566 PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKPAPTTP 624
           | ||  : |  ||| |  |    | ||  : |     :  ||| || | || |    ||
Db    1937 PTPTGTQTPT-TTPITTTTTTVTPTPTPTGTQTPTTVLITTTTTMTPTPTPTSTKSTTVTP 1995

Qy     625 KETAPTTPKEPAPTTPKKP-------APTTPETPPPTTSEVSTPTTTK-----EPTTIHK 672
             |  |    | ||  : |           |   || |||     || |       | || :
Db    1996 ITTTTTVTATPTPTGTQTPTMIPISTTTTVTPTPTPTTGSTGPPTHTSTAPIAELTTSNP 2055

Qy     673 SPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETT 732
           |: |||: |   :    |::     :|      :| |  | ||   |:     : | :|
Db    2056 PPESSTPQTSRSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPTTTSGGHTLSPPPST 2115

Qy     733 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITT---LKTTTLAPKVTTT 789
           | :|  |    |||   : |  | :| | |:|:    ||  ::|   ::|| | |  ::
Db    2116 TTSPPGTPTRGTTT--GSSSAPTPSTVQTTTTSAWTPTPTPLSTPSIIRTTGLRPYPSSV 2173

Qy     790 KKTITTTEIMNKPEE----------------------------TAKPKDRATNSKATTP 820
                  :  | |                              : |    | || :||
Db    2174 LICCVLNDTYYAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTPSPTPTPSK-STP 2232

Qy     821 KPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP 880
           | ||:  | |||   ||   |    |:             : | |  : :  :  | :
Db    2233 TPSKPSSTPSKPTPGTKPPECPDFDPPR------------QENETWWLCDCFM-ATCKY 2278

Qy     881 NQTPNSKLVEVNP-------------KSEDAGG------------AEGETPHMLL----- 910
           | |     ||  |              : ||  |              |: || :
Db    2279 NNTVEIVKVECEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGD-PHYVTFDGLY 2337

Qy     911 ------RPHVFMPEVTPDMDYLPRVPNQGIIIN 937
                 :| : |::| :|      | |: |:
Db    2338 YSYQGNCTYVLVEEISPSVD------NFGVYID 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

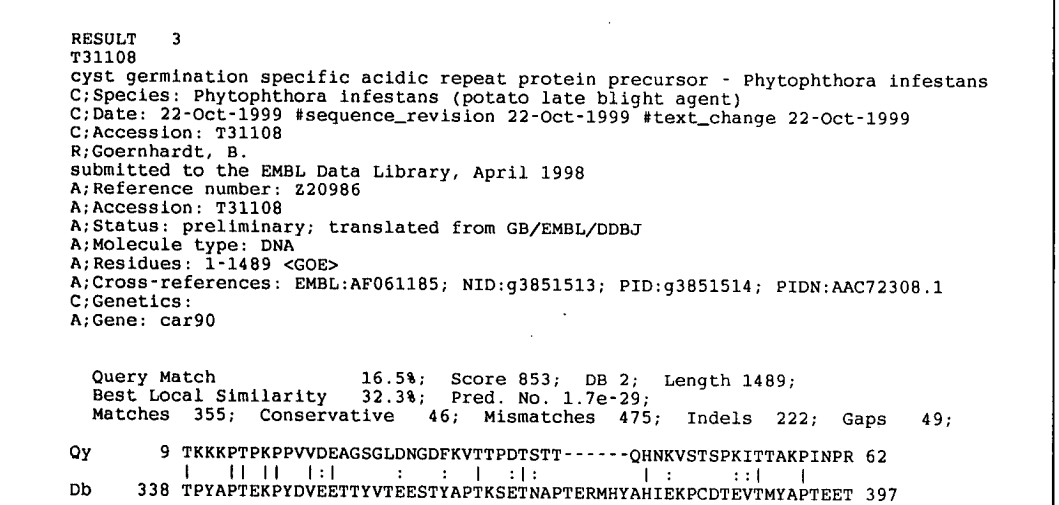
Query Match 18.3%; Score 942; DB 2; Length 1664;
Best Local Similarity 33.0%; Pred. No. 3.2e-33;
Matches 276; Conservative 83; Mismatches 312; Indels 166; Gaps 43;

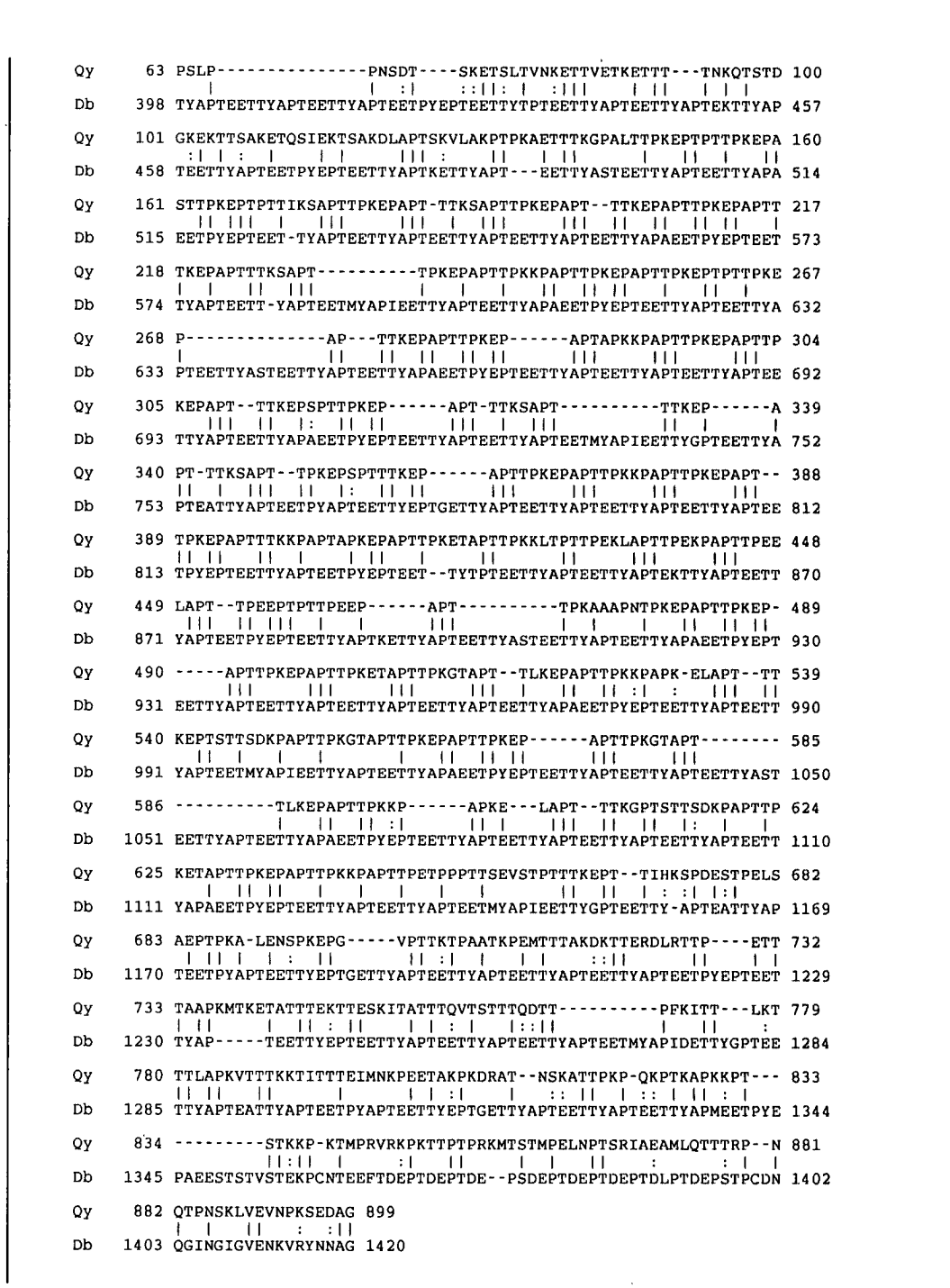
```
Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |   ||||:           ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :  ||:|| || |    ||   ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
```



RESULT 3
T31108
cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90

Query Match 16.5%; Score 853; DB 2; Length 1489;
Best Local Similarity 32.3%; Pred. No. 1.7e-29;
Matches 355; Conservative 46; Mismatches 475; Indels 222; Gaps 49;





```
RESULT   4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN00
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

  Query Match             15.6%;  Score 802;  DB 2;  Length 1274;
  Best Local Similarity   28.0%;  Pred. No. 2.2e-27;
  Matches  290;  Conservative  102;  Mismatches  414;  Indels  230;  Gaps   56;

Qy       6 KNRTKK-KP-----TP-----KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTS 50
           || ||| ||     ||     :||| ::    |   |     |   | |    || |
Db     274 KNPTKKWKPPWEDETPVEEVKEPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPS 333

Qy      51 PKITTAKPI-NPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
             :    |: ||     | :   | :  | | :  | |      |:        ::
Db     334 SPVVPPTPVKNPVKKYKPPWEVDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPS 393

Qy     110 ETQSIEKTSAKDLAPTSKVLAKPTPKAE----TTTKGPA----------------LTTP 148
             ::      :| :|      :  |||:    |  | |                  :  |
Db     394 PKKAAPAVKPRDSSPKKATPLQADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQP 453

Qy     149 KEPTPTTP----KEPASTTPKEP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAP 199
           : |   ||    ||||:    :|    || |  |  |  |:: :|    : | : : ||
Db     454 EAPAKKTPVLKRKEPAAKDTAKPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA- 512

Qy     200 TTTKEPA-------------------PTTPKEPAPTTTKEPAPT-------TTKSAPTTP 233
           |  | |                    ||  |:      |:|||             ||  |
Db     513 TPVKNPVKKWRPPWEDDETPADDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKP 572

Qy     234 KEP-----APTTPKKPAPTTPKE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EP 283
           ::|     ||| |:|  |   |:    || :  |||  : |::|:|    ||   || | |
Db     573 RDPSPKKVAPTAPEKKTPVLAKKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEV 632

Qy     284 APTAPKKPAP------TTPKEPAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPT 333
           || | ||| |       | ||:  | :|  | ||  | |      |    : ||    | |
Db     633 APAAVKKPEPISKPKDTAPKKAEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPE 691

Qy     334 TTKEPAPTTTKSAPTTP-KEPSPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPT 388
             |:  |   | ||| |  | :      |:   ||  : ||  |: |:|    | :|||
Db     692 PEKK-TPVLAKKAPTKPDSEAAADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPK 750

Qy     389 TPKEPAPTTTKKPAPTA-PKEPAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPA 442
           |  |  |   ||| | |  ::|:|   |   | :|  : | ||     :|  |   :  |
Db     751 T--EVPPAVVKKPEPVAKSRDPSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDA 805

Qy     443 PTTPEELAPTTPEEP-----TPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKE 496
           |  |  :     : |     ||  | :|:|     | |:|  :  | : |:| | : |||
Db     806 PAEPVNVPEPEKKTPVLAKKTPVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKE 865

Qy     497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTP 555
           |:|   :  :|  |    || :| |    | |   :  ||   |:|:    :|  |
Db     866 PSPKKAEPNSPVVP----PTPVKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLA 918

Qy     556 KGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPT-------TPKKPAP-----KEL 603
           |   |  ||: ||    :|   :||   |  ||||         :|||  |     :|
Db     919 K-KEPEKPKD-APKVAAKPRDPSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEA 974

Qy     604 APTTTKGPTS------TTSDKP-----APTTPKET------APTTPKEPAPTTPKKPAPT 646
             ||  | |           |:|     ||   |:|      ||  |::|   :||| ||
Db     975 PPTPVKNPVKKWKPPWEDDDEPAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPV 1031

Qy     647 TPETPPPTTSEVSTPTTTKEPTTIHKSP----DESTPELSA-EP---TPKALENSPKEPG 698
             : | |   ||  ||  | |    | |    ||:  :|| ||   ||   : :| :|
Db    1032 AAK-PDPKIPEV-PPTPVKNPVKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPA 1089

Qy     699 V-------------PTTKTPAAT------KPEMTTTAKDKTTERDLRTTPETT--TAAPK 737
                         ||:| | :      ||: ||  ||   :      ||     ||||
Db    1090 TKPDSEAAADPVSGPTSKDPKLSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPK 1149

Qy     738 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
             |          |: |         |     |           ||    |
Db    1150 KWKPVWDDDPDEPEADFTVPAPSKKPDTEDPADPLG---------GPKTKDPK------- 1193

Qy     798 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK--------- 848
           :||     || ::        |||:: :| | |||   ||   |:  ||
Db    1194 -LNKKAPAEKPTEK--------PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPE 1243

Qy     849 ---TTPTPRKMTSTMP 861
             | | |:|  :  |
Db    1244 ADFTMPAPKKPDTEDP 1259


RESULT   5
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

  Query Match             15.2%;  Score 784;  DB 2;  Length 1188;
  Best Local Similarity   29.9%;  Pred. No. 1.2e-26;
  Matches  230;  Conservative   59;  Mismatches  359;  Indels  122;  Gaps   33;

Qy      13 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 68
           |||  |  |          |:  ||          | ||::       | |||    ||
Db     460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy      69 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 128
           :                    |     ||||          |     | | |
Db     511 APVG------------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy     129 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 183
                :  | | |   :  |   :|  | |      |   :|  | || : | |   | |
Db     542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy     184 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 240
           |    ::|  |    | ||    | |  |   :|   | | | |  |:|  | | | |
Db     600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy     241 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
           | |  |   : | | |      | |   | ||    | |  |   |:|| :| |::|:||
Db     660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy     292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
           :|  ||||  : |: |       | |     | ||   | |      :| : ||:|
Db     719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy     352 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 411
             |:| ::  |||     | |     | |  ||   | |    |   |  | ||    |
```

```
Db    769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy    412 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 458
          || : |    ::|    |   |  |    | ||:  :|     : || : |:||  |
Db    828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy    459 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 517
          |||   | |: ||:: | ||   | |     | |     | | ||| : |     : | |
Db    886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy    518 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 563
          :|    ||     | |||    |||  | |   |    ||  :|  ||   || : |  |
Db    944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000

Qy    564 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 622
           | |   | | |  | : |  :| | |  |    |   | :  | : |  | |
Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    623 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 681
           :|   || :   |   :|  ||| :  :|||       |     |   |||    |
Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 731
          |  | |       | | :|    | :: | : | |  |  |: |   |:
Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158


RESULT   6
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a muscle
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor


  Query Match             15.1%;  Score 776;  DB 2;  Length 2187;
  Best Local Similarity   27.7%;  Pred. No. 4.6e-26;
  Matches  288;  Conservative  122;  Mismatches  389;  Indels  240;  Gaps   49;

Qy     34 VTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKE---------TSLTVNKETT 84
          | :|       :   |||| |:|  : |:       |||           |||:  |
Db    806 VQSPKVDPIMSDVTPTSPKKTSATAV-PK-------DTSATLSLKSVPAVTSLSPPKAPV 857

Qy     85 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPA 144
          : | |    :  | |    :|   :::  :  |  :|:     : |||: :    ||
Db    858 APSNEATIVPTEIPTSLKNALAAATPKETLATSIPKVTSPS----PQKTPKSVSLKGAPA 913

Qy    145 LTT-----------------PKE-------PTPTTPKEPASTT---------PKEPTPTT 171
          :|:                 |||       |  : || | | |         || | | |
Db    914 MTSKKATEIAASKDVSPSQFPKEVPLLQHVPPTSPPKSPVSDTLSGALTSPPPKGP-PAT 972

Qy    172 IKSAPTTPKE-PAPTTTKSAPTTPK---------EPAPTTTKEPAPTTPKEPAPTTTKEP 221
          :   || ||: | |  :|  | ||         |  | : | |    ||| : |:: :
Db    973 LAETPTYPKKSPKPAASKKTPATPSPEGVTAVPLEIPPCSKKAPKTAAPKESSATSSSKR 1032

Qy    222 APTT--TKSAPT----------------------TPKEPAPTTPKK-PAPTTPKEPAPTTP 257
          || |  :|  |:                      || | : ::||: |    |||   | |
Db   1033 APKTAVSKEIPSKGVTAVPLEISLPLKETSKSATPGEKSASSPKRSPKTAGPKE---TPP 1089

Qy    258 KEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTT--PKEPAP----TTPKEPAPTT 311
          |   |:   |  : |   || |    : :| :| |  |||  |       | |
Db   1090 GGVTAVPPEISLPPKETPQNATPNESLAASSQKRSPKTSVPKETPPGGVTAMPLEIPSAP 1149

Qy    312 TKEPSPTTPK-----EPAPTTTKSAPTTTKEPAPTTT-KSAPTTPK--------EPSPTT 357
          | |    ||      | | |    :| : |: : |    | || ||         | ||:
Db   1150 QKAPKTAVPKQIPTPEDAVTILAGSPLSPKKASKTAAPKEAPATPSVGVIAVSGEISPSP 1209

Qy    358 TKEPAPTTPKEPAPTTPKKPAPTT--PKE-----------PAPTTPKEPAP-------T 396
          |      ||| : | | | :| |  |||                |:  :|  | |       |
Db   1210 KKTSKTAAPKENSATLPPKRSPKTAAPKETPATSSEGVTAVPSEISPSPPTPASKGVPVT 1269

Qy    397 TTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK-PAPTTPEELAPTTPE 455
          | | || |  | :| :||:   |   : | |||      :|:| |    |:| : | |
Db   1270 LTPKGAPNALAE-SPASPKKVPKTAAPEETSTTP------SPQKIPKVAGPKEASATPPS 1322

Qy    456 EPTPTT--------PEEPAPTTPKAAAPNTPKEPAPTTPKE-PAPTTPKEPAPTTPKETA 506
          : || |         | |      |    |:  | |    ||| |||:   | | | | :
Db   1323 KKTPKTAVPKETSAPSEGVTAVPLEIPPSPRKAPKTAAPKETPAPS--PEGATTAPVQIP 1380

Qy    507 PTTPKGTAPTTLKEPAPTTP-------------------KKPAPKELAPTTTKEPTS-- 544
          |  ||:     ||  ||||                   |  :|||   | : :  |
Db   1381 PSPRKGSKKAGSKE-TPTTPSPEGVTAAPLEIPISSKKTSKMASPKETLVTPSSKKLSQT 1439

Qy    545 -----TTSDKPAPTTPKGTAPTTPKEPAPTTPKE-PAPTTPKGTAPTTLKEPAPTTPKKPA 599
               || : |   |    |:  | |    ||: |   :||  |||| | :|::||| |
Db   1440 VGPKETSLEGATAVPLEIPPSHKKAPKTVDPKQVPLTPSPK-DAPTTLAE-SPSSPKK-A 1496

Qy    600 PKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETP-----PPT 654
          ||  || : :  |:   :||| |  | :  |  | | |   :: | :: |||     ||
Db   1497 PKTAAPPSER-VTTVPPEKPA-TPQKASGTTASKVPVPAETQEVAVSSRETPVTPAVPPV 1554

Qy    655 TSEVSTPTTTK------EPTTIHKSPDESTPELSAEPTPKALENSPKE-PGVPTTKTPAA 707
          :  |   |:|       | |:  || :|    |::  |:   ::||| |  |: | |
Db   1555 KNPSSHKKTSKTIELKEAPATLPPSPTKSPKIPSSKKAPRT--SAPKEFPASPSIK-PVT 1611

Qy    708 TKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQ 767
          |   :  ||     :    | |:   ||| :       : |:  :   |: :   :|
Db   1612 T--SLAQTAPPSLQKAPSTTIPKENLAAPAV----LPVSSKSPAAPARASASLSPATAAP 1665

Qy    768 DTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTK 827
          | | : ||: :   |   |  :| |   :  |:||::     : ||     |
Db   1666 QTAPKEATTIPSCKKAAATETPIETSTAPSLEGAPKETSE----TSVSKVLMSSP----- 1716

Qy    828 APKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSK 887
          ||| :|:|:  |:|              :|:| |   | ::         |  | : |
Db   1717 -PKKASSSKRASTLP-------------ATTLPSLKEASVLS---------PTATSSGK 1752

Qy    888 LVEVNPKSEDAGGAEGETP 906
          ::| | ||      ||
Db   1753 DSHISPVS-DACSTGTTTP 1770


RESULT   7
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
```

```
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacchar
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseudohy
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosidase
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>

  Query Match             14.8%;  Score 763;  DB 1;  Length 1367;
  Best Local Similarity   28.5%;  Pred. No. 1.1e-25;
  Matches  286;  Conservative   98;  Mismatches  465;  Indels  154;  Gaps   43;

Qy      35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
Db     236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy      87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
Db     296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy     145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
Db     345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy     195 KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAP 254
Db     405 SSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTES 453

Qy     255 TTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEP 307
Db     454 SSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSS 509

Qy     308 APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP 363
Db     510 APVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTP 568

Qy     364 ----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETA 418
Db     569 VTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSA 624

Qy     419 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKA 472
Db     625 PVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 679

Qy     473 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAP 531
Db     680 SAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 731

Qy     532 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTL 587
Db     732 SSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 790

Qy     588 KEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKK 642
Db     791 SAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESS 850

Qy     643 PAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP 697
Db     851 SAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYP 906

Qy     698 GVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT 757
Db     907 GSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT 966

Qy     758 --TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDR 811
Db     967 VPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA----- 1021

Qy     812 ATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPR 854
Db    1022 ---SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP- 1077

Qy     855 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHV 914
Db    1078 -SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGCSPKTVTTTVPCST 1122

Qy     915 FMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTT 957
Db    1123 GTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTT 1164


RESULT   8
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of ch
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1

  Query Match             13.0%;  Score 671;  DB 2;  Length 1151;
  Best Local Similarity   25.9%;  Pred. No. 7.3e-22;
  Matches  284;  Conservative  100;  Mismatches  419;  Indels  292;  Gaps   55;

Qy      14 TPKP-------PVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPS 64
Db     113 TPPPSQGPAGTPPPSQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP- 170

Qy      65 LPPNSDTSKETSLTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDL 122
Db     171 -------ASAASPTVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EV 219

Qy     123 APTSKVLAKPTPKAETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPT 177
Db     220 KPAAAAVAKEAKAVTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--T 277

Qy     178 TPKEPAPTT-----------TKSAPTTPKEPAPTTTK---------------------- 203
Db     278 AEAKPVPATASLMATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEA 337

Qy     204 EPAPTTPKEPAPTTTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPK 250
```

```
Db    338 KPAPPPPPQQLPKAAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP- 396

Qy    251 EPAPTTPKEPTPTT----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKE 306

Db    397 -TAGAVPKASTGTTPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQ 448

Qy    307 P---APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKE 360

Db    449 PVTKAATTTNATPPPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAAT 508

Qy    361 P----APTTPKEPAPTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAP 406

Db    509 PGVPNAATDPQKPPPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIP 568

Qy    407 K--------EPAPTTPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA------ 450

Db    569 KPVPLMALTPQPVTAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPPGLPRALAAAKLLG 627

Qy    451 -PTTP----------EEPTPTTP----EEPAPTTP---------KAAAPNTPKE------ 480

Db    628 LPSSPVASAMHAKVTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNG 687

Qy    481 ----PAPTTPKEPAPTTPKEPAPTTPKETAPT------------TPKGTAPT------- 516

Db    688 TLMAPMGAANTQMAPIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGA 747

Qy    517 ------------TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTP 555

Db    748 GGATQMSPMGAANTQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSP 807

Qy    556 KG--TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP 611

Db    808 MGAATPPQTPSVGAATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPP 859

Qy    612 ----TSTTSDKP-APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP------- 652

Db    860 PMAATNTTQPPPMAASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPT 919

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKP 710

Db    920 PPPPIPPSPTAQTSPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---P 975

Qy    711 EMTTTAKDKTTERDLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDT 769

Db    976 IGVTEASPSADGARLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EA 1029

Qy    770 TPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP 829

Db   1030 AP---------TKAKRSSSSSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA--- 1077

Qy    830 KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLV 889

Db   1078 ------------PAVGDGQQQMTPGAAQSVPP-------VTEAAVQ-------------- 1104

Qy    890 EVNPKSEDAGGAEGE 904

Db   1105 EAAAAAAAAAGAERE 1119


RESULT   9
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episialin
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precursor,
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218; S51
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene enc
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129;
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burc
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epith
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated se
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky,
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may gen
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associa
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut ag
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokin
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 12
 partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48
```

```
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice form
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predicted
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted


  Query Match             12.9%;  Score 667;  DB 1;  Length 1344;
  Best Local Similarity   27.8%;  Pred. No. 1.2e-21;
  Matches  305;  Conservative   77;  Mismatches  459;  Indels  256;  Gaps   56;

Qy      77 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           |||   ||      |  |     :|| | || ||| :  |:  :: |: :: || ||:  :
Db      15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 166
           |    :||:|    |    ||               | |     ||||           :
Db      75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy     167 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 206
           | | :       : ||| |   ||| :|        ||| |   ||| :|  |       |
Db     135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 252
           | |   ||| :|  ||   | ||| |   |  | |           :||| :   |
Db     191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy     253 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 289
             || |   |  | |            ||| :  |       || |   |  |||
Db     248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy     290 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 342
             || |   |  | |       || |   |  | |     |:   |: || :|  ||
Db     308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy     343 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 391
           | ||| |   |:| :|  |       || |   |  | |       || |   |  | |
Db     368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy     392 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 447
               || |   |  |||     |:  :| |       |      || |   |  | |
Db     425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy     448 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 487
               || |   |  | |            |||  | | |    :||:|   |  | |
Db     485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy     488 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 534
               || |   |  | |     :|| |   |  | |           :|||
Db     545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy     535 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 588
           ||      ||    :|||  | |      |:    || |   |  | |      :|| |
Db     601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy     589 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 632
             | | |           :|||   ||    | ||    :|||  | |      :|| |
Db     656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy     633 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 675
                 |  | |           :||| :  | ||     | | |   | :       : :||
Db     715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy     676 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 724
               |  |    : ::|     | :  | |         |:|   :||      |:  |
Db     773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy     725 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 775
           |  | :|       |:||    :|      |       |:   |     | |    :|
Db     833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy     776 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 833
           :   |  ||  |       |:  :    |  || |    |::  | | |   : ||
Db     889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy     834 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 885
           |  | | |    | :|  |    :: |:  |    |:  |  :     ||| :  |
Db     947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy     886 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 945
           :  |   | :  | |:     |           ||   |  | : :      | : : |:
Db    1004 AHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy     946 ICNGKPVDGLTTLRNGT 962
                 | : |||
Db    1050 ASGSASGSASTLVHNGT 1066


RESULT  10
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3


  Query Match             12.8%;  Score 659.5;  DB 2;  Length 1229;
  Best Local Similarity   29.0%;  Pred. No. 2.4e-21;
  Matches  307;  Conservative   78;  Mismatches  339;  Indels  333;  Gaps   52;

Qy      33 KVTT-PDTSTTQ---HNKVSTSPKITTA-KPINPRPSLPPNSDT-------SKETSLTVN 80
           ||||  | |||       | ||:|:| |    ||  :     :  |       | | | :
Db     119 KVTTSTDASTTNAPTTGKDSTTPEIITGIVVINSKSESVTDMSTTRFSTTLSPTTELLTS 178

Qy      81 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 140
           || | |  :|:| :||| |   :  |  || :  : :   : |:   ||    : |||
Db     179 PETLVSTDSSTST-EQTSPDNTTEIASPMETNTTTEATTTSVEPSVSTLAS---EDETTV 234

Qy     141 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 200
             | :|      ||     ||| :||| |            | :||| : |     ||
Db     235 TAIAEST------TTVIAEVSTTTEEPTTT-----------AESTTKKSTT----KAPA 272

Qy     201 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----KKPAPTTPKEPAPTT 256
           ||:|| ||| :|   |||:    ||| |  :| |   ||||    |   | | |   | |
Db     273 TTEEPTPTTTEE--VTTTEAETSTTTSSETSTEK---PTTPLIDNKIAGPATGK---PET 324

Qy     257 PKEP-TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKE- 314
             | | |||     | |       |   | | |  | | ::    |   |  |||
Db     325 THFPVTGTTPN--FDTATETPFVAKSEDKMTLSKTAATETTQQTTEVT-DGPEKETTKNV 381
```

```
Qy    315 ----PSPTTPKEPAPTTTKSAP-------TTTKEPAPT----TTKSAPTTPKEPSPTTTK 359
Db    382 SIEIPITTVPLVETTSTTSTASKESDGFHTTLKLKVTTADSDSTESATTVKPFNEETTTK 441

Qy    360 E---PAPT-------TPK------EPAPTTPKKPAP--------------------TTPK 383
Db    442 SHVVPKPTKKGTVKVTPKLELSFDEPTEIT-KAPHPGKLLEKKTYHFVLSDNFARYSEAK 500

Qy    384 E---------------PAPTTPKEPAP----TTTKKPAPTA--PKEPAPTT---PKETAP 419
Db    501 ENDDYNHLDYNHYREAKEPTTTEESSTTEEVTTTEEPANTGNPPTTENPTTTEQPTSTAE 560

Qy    420 TTPKKLTPTTPEKLA---PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPN 476
Db    561 STTTALPFTTEQTVTTEEPTTAEKSTATQ----KPTTTQESVST---EKTSTTKKA---S 610

Qy    477 TPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 536
Db    611 TTEE--PTTTDEPTTTT---ESSTTGKATTPELSTTSEETTTTELKITTE--------GS 657

Qy    537 TTTKEPTSTTSDKPAP----TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAP 592
Db    658 TTTEEPTTTAIFAEASTGIITTDEETTSTTSTTPEITSTKE--IVTESAITQTSVSVVES 715

Qy    593 TTP----------------------KKPAPKELAPTTTKGP---------------TS 613
Db    716 STPRQLPERWKAIVNKFKHNLEVLKEKKRLLKEKESTSTTGSDSSETTTVVAENIDEVTT 775

Qy    614 TTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPE--TPPPTTSE-VSTPTTTKEPTTI 670
Db    776 TEKEKVVQTTPITTEKSTTQEETTTTTTTTEKTTSKTTTEKPTTSESATTETTTSEPST- 834

Qy    671 HKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPE 730
Db    835 ----TEST----------------------TVDTSSATTEESSTAAETTTTSAE---TSE 865

Qy    731 TTTA----------------------------------------APKMTKETATTTEKTTE 751
Db    866 TTTSESAAFITGESPENTALQSSSQKSEENESSAEKPGARRDFVPKKHKTTVKPAETTSA 925

Qy    752 SKITATTTQ-VTST---TTQDTTPFKITTLKTTTLAPKVT---TTKKTITTTEIMNKPEE 804
Db    926 VAASTTTTEPITTTEKSTTLETTPIEATTLNEVTGPAFVTGAPVDETTINTLELLSK--- 982

Qy    805 TAKPKDRATNSKATTPKPQ----------------KPTKAPKKPTSTKKPKTMPRVRKP 847
Db    983 -------INNTQISQPKPTDISKTDALSSLISGLIGSFTKAPMAPTI------------ 1022

Qy    848 KTTPTPRKMTSTMPELN------------PTSRIAEA 872
Db   1023 HTTTDAAFVTATEASLNDGSDKKIIDEAQPTDEIRRA 1059
```

RESULT 11
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume inte
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>

Query Match 12.3%; Score 633; DB 2; Length 489;
Best Local Similarity 31.2%; Pred. No. 1.4e-20;
Matches 149; Conservative 33; Mismatches 252; Indels 44; Gaps 7;

```
Qy    179 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 238
Db     39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy    239 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 295
Db     84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy    296 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 355
Db    142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP--------------PPSPSPPPP 187

Qy    356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 415
Db    188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy    416 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 475
Db    248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy    476 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 535
Db    308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYY- 366

Qy    536 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 595
Db    367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy    596 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 653
Db    423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPP 475
```

RESULT 12
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31

Query Match 12.3%; Score 632.5; DB 2; Length 7962;

```
  Best Local Similarity   22.9%;  Pred. No. 1.9e-19;
  Matches  295;  Conservative  105;  Mismatches  453;  Indels  437;  Gaps   60;

Qy      10 KKKPTP----KPPV---VDEAGSGLDNGDFKVT-----------TPDTSTTQHNKVSTS 50
Db    5837 KKAPPPPKVIKKPVIEKIEKTSRRMEEEKVQVTKVPEVSKKIVPQKPSRTPVQEEVIEVK 5896

Qy      51 -PKITTAKPINPRPSLPPNSDTSKETSLT---VNKETTVETKETTTTNKQTSTDGKEKTT 106
Db    5897 VPAVHTKKMVISEEKMFFASHTEEEVSVTVPEVQKEIVTEEKIHVAVSKRVE-------P 5949

Qy     107 SAKETQSIEKTSAKDLAPT--SKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPAS--- 161
Db    5950 PPKVPELPEKPAPEEVAPVPIPKKVEPPAPKVPEVPKKP-VPEEKKPVPVPKKEPAAPPK 6008

Qy     162 --TTPKEPTPTTIKSAPTTPKEPAP-----------TTTKSAPTTPKEPAPTTTKEPA- 206
Db    6009 VPEVPKKPVPEEKIPVPVAKKKEAPPAKVPEVQKGVVTEEKITIVTQREESP----PPAV 6064

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP-KEPAPTTPKEPTPTT- 264
Db    6065 PEIPKKKVPEERK-PVPRKEEEVPPPPKVPA--LPKKPVPEEKVAVPVPVAKKAPPPRAE 6121

Qy     265 ----------------------------------------------------------P 265
Db    6122 VSKKTVVEEKRFVAEEKLSFAVPQRVEVTRHEVSAEEEWSYSEEEEGVSISVYREEEREE 6181

Qy     266 KEPAPTTK-------------------------EPAPTT----------PKEPA------ 284
Db    6182 EEEAEVTEYEVMEEPEEYVVEEKLHIISKRVEAEPAEVTERQEKKIVLKPKIPAKIEEPP 6241

Qy     285 ----PTAPKK-------PAPTTPKE--PAPTTPKEP-APTTTKEPSPTTPKE----PAPT 326
Db    6242 PAKVPEAPKKIVPEKKVPAPVPKKEKVPPPKVPEEPKKPVPEKKVPPKVIKMEEPLPAKV 6301

Qy     327 TTKSAPTTTKEP--APTTTKSAP---TTPKEPSPTTTKE-------------PAPTT--- 365
Db    6302 TEKHMQITQEEKVLVAVTKKEAPPKARVPEEPKRAVPEEKVLKLKPKREEEPPAKVTEFR 6361

Qy     366 -------------PK-EPAP---------------------------------------- 371
Db    6362 KRVVKEEKVSIEAPKREPQPIKEVTIMEEKERAYTLEEEAVSVQREEEYEEYEEYDYKEF 6421

Qy     372 ------------------------------TTPKKPAPT--TPKEPAPTTPKEPAPTTTK 399
Db    6422 EEYEPTEEYDQYEEYEEREYERYEEHEEYITEPEKPIPVKPVPEEPVPTKPKAPPAKVLK 6481

Qy     400 KPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK-------LAPTTPEKPAPTTPEELAPT 452
Db    6482 KAVP-EEKVPVP-IPKKLKPPPPK--VPEEPKKVFEEKIHISITKREKEQVTEP---AAK 6534

Qy     453 TPEEPTPTTPEEPAPTTPKAAAP--NTPKEPAPTTPKEPA---------------------- 490
Db    6535 VPMKPKRVVAEEKVPVPRKEVAPPVRVPEVPKELEPEEVAFEEEVVTHVEEYLVEEEEEY 6594

Qy     491 ------PTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTS 544
Db    6595 IHEEEEFITEEEVVPVIPVK-VPEVPRKPVPEE-KKPVPVPKKKEAPPAKVPEVPKKPEE 6652

Qy     545 TT------SDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPK 596
Db    6653 KVPVLIPKKEKP----PPAKVPEVPKKPVP-EEKVPVPVPKKVEAP-----PAKVPEVPK 6702

Qy     597 KPAPKEL----APTTTKGPTSTTSDKPAPTTPKETAPT-TPKE---PAPTTPKKPAPTTP 648
Db    6703 KPVPEKKVPVPAPKKVEAPPAKVPEVPKKLIPEEKKPTPVPKKVEAPPPKVPKKREPVPV 6762

Qy     649 ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAAT 708
Db    6763 PVALPQEEEVLFEEEIVPEEEVLPEEEEVLPE-EEEVLPEEEEVLPEEEEIPPEEEEVPP 6821

Qy     709 KPEMTTTAKDKTTERDL--RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
Db    6822 EEEYVPEEEEFVPEEEVLPEVKPKVPVPAP-----VPEIKKKVTEKKVVIPKKEEAPPAK 6876

Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPT 826
Db    6877 VPEVPKKVE--EKRIILPK----EEEVLPVEVTEEPEE----------------EPISEE 6914

Qy     827 KAPKKPTSTKK-----PKTMPRVRK---PKT-TPTPRKMTSTMPELNPTSRIAEAMLQTT 877
Db    6915 EIPEEPPSIEEVEEVAPPRVPEVIKKAVPEAPTPVPKKVEA--PPAKVSKKIPEEKVPVP 6972

Qy     878 TRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIIN 937
Db    6973 VQKKEAPPAKVPEV-------------------------PKKVPEKKVL--VPKKEAV-- 7003

Qy     938 PMLSDETNICNGKPVDGLTTLRNGTLVAFR 967
Db    7004 ------------PPAKGRTVLEEKVSVAFR 7021


RESULT  13
C84672
hypothetical protein At2g27380 [imported] - Arabidopsis thaliana
C;Species: Arabidopsis thaliana (mouse-ear cress)
C;Date: 02-Feb-2001 #sequence_revision 02-Feb-2001 #text_change 02-Feb-2001
C;Accession: C84672
R;Lin, X.; Kaul, S.; Rounsley, S.D.; Shea, T.P.; Benito, M.I.; Town, C.D.; Fujii, C.Y
M.; Koo, H.; Moffat, K.S.; Cronin, L.A.; Shen, M.; VanAken, S.E.; Umayam, L.; Tallon,
euss, D.; Nierman, W.C.; White, O.; Eisen, J.A.; Salzberg, S.L.; Fraser, C.M.; Venter
Nature 402, 761-768, 1999
A;Title: Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana.
A;Reference number: A84420; MUID:20083487
A;Accession: C84672
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-761 <STO>
A;Cross-references: GB:AE002093; NID:g5306260; PIDN:AAD41992.1; GSPDB:GN00139
C;Genetics:
A;Gene: At2g27380
A;Map position: 2


  Query Match           12.3%;  Score 632;  DB 2;  Length 761;
  Best Local Similarity   31.1%;  Pred. No. 2.3e-20;
  Matches  217;  Conservative   44;  Mismatches  354;  Indels   82;  Gaps   33;

Qy     132 PTPKAETTTKGPALTTP---KEPTPT-TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTT 187
Db      69 PPPIQKPPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPT 128

Qy     188 KSAPTTPKEPAPTTTKEPAP--TTPKEPAPTTTKEPAPTTTKSAPTTP----KEPAPT-- 239
Db     129 YSPPIYP----PPIQKPPTPSYSPPVKPPPV---QMPPTPTYSPPIKPPPVHKPPTPTYS 181

Qy     240 TPKKP---APTTPKEPAPTTP----KEPTP--TTPKEPAPTTKEPAPT-TPKEPAPTAPK 289
Db     182 PPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHK 241

Qy     290 KPAPT-TPKEPAPTTPKEPAP---TTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 345
Db     242 PPTPIYSPPIKPPPVHKPPTPIYSPPVKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPP 301

Qy     346 APTTPKEPSPTTTKEP---APTTPKEPAPT--TPKKPAPTTPKEPAPTTPKEPAPTTTKK 400
Db     302 VKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPVKP-PPVHKP 360

Qy     401 PAP--TAPKEPAPT----TPKETAPTTP----KKLTPTTPEKLAPTTPEKPAPTT---PE 447
Db     361 PTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPI 420

Qy     448 ELAPTTPEEPTPTTPEEPAPT----TPKAAAP-NTPKEPAPTTPKEPAPTTPKEPAPTTP 502
Db     421 KLPPVKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTP 480
```

```
Qy    503 KETAPTTPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
Db    481 TYSPPVQP----PPVQKPPTPTYSPPVKPPPIQKPPT----PTYSPPIKPPPVKPPTPTY 532

Qy    561 TTPKEPAPTTPKEPAPT-TPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKGPTSTTSD 617
Db    533 SPPIKP-PPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPP 591

Qy    618 KPAPTTPKETAPTTPKEPAPTTPKKPAPT-TPETPPPTTSEVSTPTTTK--EPTTIHKSP 674
Db    592 VHKPPTPTYSPPIKP----PPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPP 647

Qy    675 DESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA 734
Db    648 ---TPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPP-TPTYSPPVKPPPVQVPPTPTYS 703

Qy    735 APKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP 771
Db    704 PPVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPP 740
```

RESULT 14
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/2; 3504/1

Query Match 12.2%; Score 629; DB 2; Length 3507;
Best Local Similarity 24.1%; Pred. No. 1.2e-19;
Matches 252; Conservative 158; Mismatches 416; Indels 218; Gaps 45;

```
Qy     10 KKKPTPKPPVVDEAGSGLDN-GDFKVTT----PDTSTTQHNKVSTSPKITTA-------K 57
Db   1942 KKQPNREKIEIDEENSSSSNSGQEKPTTKGIVSSTSATSSESTTAEPHVTTSISSTTSTK 2001

Qy     58 PINPRPSLPPN---SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQS 113
Db   2002 DMTSSKS-PENVTMSSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEV 2060

Qy    114 IEKTSAKDLAP----TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
Db   2061 ITESSVKSTTPKEESSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKS 2119

Qy    170 TTIKS-APTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKS 228
Db   2120 TVLSSEAPVTSTSPTEVHT-SSETKPSLSASSTTGDTNSTTPSTSSLASVKSTSAPEGTS 2178

Qy    229 APTTPKEPAPTTPKKPAPTTPK-EPAPTTPKEPTPTTPKEPAPTTKEP-APTTPKEPAPT 286
Db   2179 ASVAPVKLSSLSPDVSQPSTKTFDATESSTVQASETSSGTSVKSTSEPESHVTKLSITSS 2238

Qy    287 APKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP--KEPAPTTTKSAPTTTKEPAPTTTK 344
Db   2239 NPSSSVPVTSPKSTPTVPE-----STEQPTSTTPSGQSLTPMNSNSEVLTTSEPHVLSSS 2293

Qy    345 SAP------TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT- 397
Db   2294 LSPDVSQSSTTPNNLSESSTVE----TPKTSSEVSLNSEEPSTTEAPTTLSPDILSTTTN 2349

Qy    398 --------------------TKKPAPTAPK----------EPAPTTPKETAPTTPKKLTP 427
Db   2350 NLSQSSTVSTEDRSEISSENSEKPT-SAPELVTSSVTHVASSSPDVPTES--SEPDDLTG 2406

Qy    428 TTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP-----------APTTPKAAAPN 476
Db   2407 SSTENIPEASSKQTISSTPTPDTTTASEEPTKSTSMSPDLSTTSNVLSESSTTPESSS-- 2464

Qy    477 TPKEPAPTT------------PKEPAPT---------TPKEPAP-----TTPKETAPTTP 510
Db   2465 --KSPVSSSTEGISVVTSTEFSKVPESTISSVLEEDLTKTTPSPILEETTTASETSEPLT 2522

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPA---------PT-------- 553
Db   2523 EDSLTVSVRIHELTTSSENVPKESESTTTSSESSKPSQEPAGILTSTVVVPTSSVSLITA 2582

Qy    554 ---------TP--KGTAP--TTPK-------EPAPTTPKEPAPTTPKGTAPTTLKEPAPT 593
Db   2583 SEIEAITSNTPFKQGRTPITTSPKSLVKSTTSPSTVTSSEPSESTKRTTVSTTVSTTTPT 2642

Qy    594 TPKKPAPKELAPTTTKGPT-STTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPP 652
Db   2643 EETTTSESLILTAAPSKPTESTTESSEAPTTPAKTSETKPSNVSSTSRKSTENVETSTSQ 2702

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTP--ELSAEPTPKALE--NSPKEPGVPTTKTPAAT 708
Db   2703 SGSLESSTMSSTS------SEPETNAPAVTVSSEASSTTLEENSSTSSP----TSSEASV 2752

Qy    709 K-----PEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTT---- 759
Db   2753 KLSSLFPESITSEAVTVSSR----APAEITMSSESHREISTVSSEPSEPEIPLSTTVSPN 2808

Qy    760 -----------QVTSTTTQDTTP-FKITT-----LKTTTLAPKVTTTKKTITTTEIMNKP 802
Db   2809 VVTASSIPSEEPILSSVTSSSTPRVRLITGTPDDLIVSVTVPSHGNRRQNITASSV---P 2865

Qy    803 EETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK---TTPTPRKMTST 859
Db   2866 SNSTSPIILPSES-LTTPQPPPTTTTTAKPATTSGKRGPPSIQPPAEMFTTPAP------ 2918

Qy    860 MPELN-----PTSRIAEAMLQTTT 878
Db   2919 PPPSNGGYGEETNQEEEQVTSTTT 2942
```

RESULT 15
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp
A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNopp180
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein

Query Match 12.2%; Score 626.5; DB 2; Length 990;
Best Local Similarity 26.2%; Pred. No. 4.9e-20;
Matches 253; Conservative 132; Mismatches 393; Indels 187; Gaps 45;

```
Qy       3 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 62
Db      59 DAKKRKPPANGLPKKKSAKESSSSEDSSSEEDEPPAKKRAQ----PAGGKKPVVKAVQPK 114

Qy      63 PSLPPNSDTSKETSLTVNKETTVE--TKETTTTNK----QTSTDGKEKTTSAKETQSIEK 116
Db     115 KAKSSSEDSSDESD---SEEETKKPPAKRPAQTPKVAAVKTPTQKKAKSSSS-ESSSSED 170

Qy     117 TSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKEPTPTTI 172
Db     171 EASKKKQPVIKV----PPKQAVVKAGLASNNGKTADSSSSEDSDSPPAKKTAATKTPPT- 225

Qy     173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTT-TKEPAPTT-----T 226
Db     226 --KPATAAKPQAKKTAGKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVPPPTSVSKKKT 283

Qy     227 KSAPTTPKEPAPT----------TPKKPAPTTPKEPAPTTP----KEPTPTTPKEPAPTT 272
Db     284 LSQPGTKAKPESSDSSDSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSESDSSS 343

Qy     273 KEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTP----KEPAP-----------------TT 311
Db     344 EDEKKSSVKLGVKAAPKK-APAAP--DAKSTPVAAAKKSAPAKKASSSSDSDSSSNEETT 400

Qy     312 TKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA- 370
Db     401 TKPAAKTTPAKSAATPTSKTPTNGK---ATPTSKTPAKPGTPKTSTAKKDSSSSDSSDSS 457

Qy     371 ---PTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPA-PTTPKETA------- 418
Db     458 SDEETTTKPAAKTTPAKSAATPTSKTP---TNSKATPTSKKTPAKPGTPKTSAAKKDSSS 514

Qy     419 -----PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP-KA 472
Db     515 SDSSDSSSDEKKTPA--KRAAKTTPAKPA-------AKTTPAKPA-------AKTTPAKP 558

Qy     473 AAPNTPKEPAPT----TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKEPAPTTPK 527
Db     559 AAKSTPGKQVPTKKESSSSDSSDSSSEDEKKSSAKPAVKTTPGKATS-------KPVVAS 611

Qy     528 KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT--PKEPAPTTPKGTAPT 585
Db     612 KPVPAK------KASSSSDSDSSEEETTKTTKPLTKLSPAVKTLPPKKAESSSDSSSDSD 665

Qy     586 TLKEPAPTTPKKPAPKELAPTTTKGP--------TSTTSDKPAPTTPKETAPT--TPKEP 635
Db     666 SEKK---TKPAKPPAKSATPVNTKAPAQNKASKASCSDSDSSSEEEGKSKQPTGKSPAAK 722

Qy     636 APTTPKK-PAPTTPETPPPTTSEVSTPTTTKE-PTTIHKSPDESTPELSAEPTPKALENS 693
Db     723 ATAPPKKNPVAVNKDKPSSSSSSDSSGDDEKQKPKQAAAAKDVKQGAKAAKPTPKKAASS 782

Qy     694 PKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESK 753
Db     783 SSE---DSSSDEDVSKAKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGG 835

Qy     754 ITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKD 810
Db     836 KNTSTTKIANST-----------------PKAAAAECSEESSSSEDEGKANGTSGKRKR 877

Qy     811 RATNS---KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPT 866
Db     878 ESTGNAECEAVTPE--------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP- 928

Qy     867 SRIAE 871
Db     929 -RMAD 932
```

Search completed: April 26, 2002, 16:18:59
Job time: 329 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:19:21 ; Search time 47.4 Seconds
(without alignments)
748.767 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1167
Perfect score: 5155
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....GKPVDGLTTLRNGTLVAFRG 968

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1168	22.7	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	942	18.3	1664	1	SLP1_CLOTM	Q06852 clostridium
3	763	14.8	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	12.6	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	556.5	10.8	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	10.8	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	10.7	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	10.3	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	10.3	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	10.3	865	1	CPN_DROME	Q02910 drosophila
11	518.5	10.1	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	9.9	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	9.8	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	9.7	1970	1	RPB1_MOUSE	P08775 mus musculu
15	497	9.6	831	1	NFH_RAT	P16884 rattus norv
16	493.5	9.6	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	9.5	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488	9.5	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	9.2	267	1	EXTN_MAIZE	P14918 zea mays (m
20	471.5	9.1	5376	1	ZAN_MOUSE	O88799 mus musculu
21	468.5	9.1	634	1	HWP1_CANAL	P46593 candida alb
22	467	9.1	817	1	VRP1_YEAST	P37370 saccharomyc
23	454	8.8	797	1	VGLX_HSVEB	P28968 equine herp
24	454	8.8	1161	1	YJ9P_YEAST	P47179 saccharomyc
25	453.5	8.8	2142	1	BAT2_HUMAN	P48634 homo sapien
26	448.5	8.7	670	1	VG50_HSVI1	Q00130 ictalurid h
27	442.5	8.6	751	1	FP1_MYTGA	Q27409 mytilus gal
28	439.5	8.5	1083	1	T2D3_HUMAN	O00268 homo sapien
29	432.5	8.4	439	1	XP2_XENLA	P17437 xenopus lae
30	432	8.4	3164	1	TEGU_HSV11	P10220 herpes simp
31	426	8.3	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
32	424.5	8.2	1125	1	MAP4_MOUSE	P27546 mus musculu
33	424	8.2	307	1	SGS3_DROME	P02840 drosophila
34	424	8.2	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
35	422.5	8.2	1794	1	YAV1_SCHPO	Q10172 schizosacch
36	422.5	8.2	3421	1	TEGU_HSVEB	P28955 equine herp
37	419.5	8.1	2476	1	ZAN_PIG	Q28983 sus scrofa
38	419	8.1	1185	1	DRPL_HUMAN	P54259 homo sapien
39	416.5	8.1	2774	1	MAPA_RAT	P34926 rattus norv
40	415	8.1	1229	1	N121_HUMAN	Q9y2n3 homo sapien
41	411.5	8.0	907	1	VGP3_EBV	P03200 epstein-bar
42	410	8.0	1183	1	DRPL_RAT	P54258 rattus norv
43	407.5	7.9	3256	1	KI67_HUMAN	P46013 homo sapien
44	405.5	7.9	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
45	405	7.9	1411	1	TCOF_HUMAN	Q13428 homo sapien

ALIGNMENTS

RESULT 1
MUC2_HUMAN
ID MUC2_HUMAN STANDARD; PRT; 5179 AA.
AC Q02817; Q14878;
DT 01-JUN-1994 (Rel. 29, Created)
DT 01-NOV-1997 (Rel. 35, Last sequence update)
DT 20-AUG-2001 (Rel. 40, Last annotation update)
DE MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN MUC2 OR SMUC.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Intestine;
RX MEDLINE=94132002; PubMed=8300571;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT Identification of the amino terminus and overall sequence similarity
RT to prepro-von Willebrand factor.";
RL J. Biol. Chem. 269:2440-2446(1994).
RN [2]
RP SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC TISSUE=Colon;
RX MEDLINE=93016075; PubMed=1400449;
RA Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA Kim Y.S.;
RT "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT both upstream and downstream of its central repetitive region.";
RL J. Biol. Chem. 267:21375-21383(1992).
RN [3]
RP SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX MEDLINE=91358717; PubMed=1885763;
RA Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA Petersen G.M., Kim Y.S.;
RT "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT and polymorphism.";
RL J. Clin. Invest. 88:1005-1013(1991).
CC -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC AGENTS AT MUCOSAL SURFACES.
CC -!- SUBUNIT: MULTIMERIC.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC BRONCHUS, CERVIX AND GALL BLADDER.
CC -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC VARIES AMONG DIFFERENT ALLELES.
CC -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC OF SILKWORM HEMOCYTIN.
CC -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC ---

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;
```

```
  Query Match             22.7%;  Score 1168;  DB 1;  Length 5179;
  Best Local Similarity   32.5%;  Pred. No. 4.8e-39;
  Matches  363;  Conservative   53;  Mismatches  448;  Indels  254;  Gaps   40;

Qy      21 DEAGSGLDNGDFK----------------VTTPDTSTTQH-NKVSTSPK----------- 52
           |   || |:|| :                |  |  |  ||  ||
Db    1312 DHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQF 1371

Qy      53 ------------------------ITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
                                    |||  |    || || : |:   : | :  ||
Db    1372 GNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTT---- 1427

Qy      89 ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTP 148
           ||||   |:|     ||:               ||:     |:|   |||  |  |||
Db    1428 -TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPISTTTTPPPTTTP 1470

Qy     149 KEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------APTTTKSAP-TTP 194
           ||      |||   | :||   | |||  | | |||  |         ||||  | | ||
Db    1471 SPPTTTPSPPTTTPSPPTTTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTT 1530

Qy     195 KEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPKEPAPTTPKKPA 245
             | ||||  |  |||  |        ||||  | |||| : |  |||  |  |||  |
Db    1531 TTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPT 1590

Qy     246 PTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAPKKPAPTTPKE 298
           ||   |  |||  || ||   |  ||  |  |||  |         ||    | |||
Db    1591 ITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTT 1650

Qy     299 PAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP---TTPKEPSP 355
           | |||   |  |||  |  ||   | |||| |:| ||    |||| : |   |||  |
Db    1651 PPPTTTPSPPTTTTPSPPITTTTPPPTTTPSSPITTTPSPPTTTMTTPSPTTTPSSPI- 1709

Qy     356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTP- 414
           |||  |: ||   | |||   |:|||   |  ||     |||| |  | |  |: | |
Db    1710 TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPT 1769

Qy     415 ----KETAPTTP-------------------------------------------KKLTP 427
               | ||||                                            : |
Db    1770 FSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYP 1829

Qy     428 ------------------------------------------------TTPEKLAPTT 437
                                                           | |  :  ||
Db    1830 DVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTMTTTT 1889

Qy     438 PEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKEP-----AP 491
           | | | |   :  ||   |||| |  :   |||          | ||  : |
Db    1890 TENPTPPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT 1949
```

```
Qy     492 TTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKP 550
           ||   | || |  :|  |||  |  |    | ||  : |    :  |||  || | :
Db    1950 TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQ 2009

Qy     551 AP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 609
            | |||  |  |    | ||  : |  |||  |  |    | ||  : |    :  |||
Db    2010 TPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV 2068

Qy     610 GPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT 668
            || | :    | |||  |  |    | ||  : |  |||  |  : |   ||| |: ||
Db    2069 TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPT 2127

Qy     669 TIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPAATKPEMTTTAK 717
           |   :|  :|  ::  |||   :     |             || |:||  |   |||
Db    2128 T---TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2184

Qy     718 DKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTTQVTSTTTQDT 769
           |    :   |||  |||  |  |    :| |||  || : :| | |  | | |  |
Db    2185 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP-TGTQTPTT 2243

Qy     770 TPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKA 828
           ||  |||  |||: |  | |  :| ||| |      |  |    | :  |||     |
Db    2244 TP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2299

Qy     829 P-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAMLQTTTR-PNQT 883
           |   || |: | | |       ||||    :  |   |   |: :      | |: | |
Db    2300 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT 2359

Qy     884 PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 921
           | :    | |     |    :||     |     |||
Db    2360 PITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392


RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL        1     28       POTENTIAL.
FT   CHAIN        29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN       36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT       36    191       1.
FT   REPEAT      207    363       2.
FT   REPEAT      409    565       3.
FT   REPEAT      607    763       4.
FT   DOMAIN      771   1377       APPROXIMATE TANDEM REPEATS OF
FT                                T-P-S-D-E-P.
FT   DOMAIN     1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN     1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN     1495   1565       SLH 2.
FT   DOMAIN     1566   1625       SLH 3.
FT   DOMAIN     1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;


  Query Match             18.3%;  Score 942;  DB 1;  Length 1664;
  Best Local Similarity   33.0%;  Pred. No. 1.3e-30;
  Matches  276;  Conservative   83;  Mismatches  312;  Indels  166;  Gaps   43;

Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |   ||||:           ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :   ||:|| || |    ||   ||      || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     302 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPTTTKSAPTTPKEPSPTT 357
           :    |:  || : ||:|: ||:|| || | |  || : || |:          ||
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS----------DEPTP 968

Qy     358 TKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPAPTTPKE 416
           : || |:   || |:      ||   || |: ||:|| || |    ||   || |:
Db     969 SDEPTPS--DEPTPS----DEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSD--- 1019

Qy     417 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAP 475
                 ||  ::  ||  ::| |:      ||  :|||| |||| ||   :  |
Db    1020 ---------EPTPSDE--PTPSDEPTPSD----EPTPSDEPTPSETPEEPIPTDTPSDEP 1064

Qy     476 NTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK 532
           || |:   |:  ||   || |: ||:|  ||      ||   || |:   :| |
Db    1065 TPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS 1122

Qy     533 ELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA 591
           :  || : ||| | | ::| ||      ||   || |:   || |:      ||   ||
Db    1123 D-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS----DEPTPSDEPT 1175

Qy     592 PT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKEPAPT-TPKKPAPT 646
           |: ||::|      || |     | ||:| |:   || :  ||   || |: ||::| ||
Db    1176 PSETPEEP-----IPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETPEEPIPT 1229

Qy     647 TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPA 706
           :  || |:   || : |||      || ||  | ||||     :|:|| :||
Db    1230 DTPSDEPTPSD--EPTPSDEPT----PSDEPTP--SDEPTP---SETPEEP-IPTDTPSD 1277

Qy     707 ATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
             |   |  |: |  |  |  :  | : : |     |  :| |  |   |     | :
Db    1278 EPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPS 1337

Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ-KP 825
           : ||   |                  | :   | |  | |   ||  :  |
Db    1338 DEPTPSDEPT-----------------------PSDEPTPSDEPTPSDEPTPSDEPTP 1372

Qy     826 TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTS--------------TMPELNPTSRIA 870
           :: |::|| |  |        | ||||      |              | |   |||
Db    1373 SETPEEPTPTTTP-----TPTPSTTPTSGSGGSGGSGGGGGGGGGGTVPTSPTPTPTS---. 1424
```

```
Qy    871 EAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLP 927
                 :|  ||   :|  |   |  || |      || :       ||  : |
Db   1425 --------KPTSTPAPTEIE-EPTPSDVPGAIGGEHRAYLRGY-------PDGSFRP 1465


RESULT  3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;


  Query Match             14.8%;  Score 763;  DB 1;  Length 1367;
  Best Local Similarity   28.5%;  Pred. No. 1.1e-23;
  Matches  286;  Conservative   98;  Mismatches  465;  Indels  154;  Gaps   43;

Qy     35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
          :| ::|||  :   :|   :|  |  |         :|: |  :  :||     : :||
Db    236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy     87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
          ||: |||:| | |   :|||:   | |   |:    ||          ||| : ||  :  |
Db    296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy    145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
           ::  |      |||::    :|: |   : |  |||    ||    ||  | |: ||
Db    345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy    195 KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAP 254
             || |:     ||    || |:   : | : ||| |    : ||    || |
Db    405 SSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTES 453

Qy    255 TTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEP 307
          ::   |||   ||    || |   : ||    ||  |   : ||    ||    ||
Db    454 SSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSS 509

Qy    308 APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP 363
          ||  |   | ||    ||  | |: ||    ||    ||  |:   |   | ||     |
Db    510 APVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTP 568

Qy    364 ----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETA 418
              ||    ||  ||      :   | ||    |: :||:  :  ||    : ||   :|
Db    569 VTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSA 624

Qy    419 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKA 472
          | |    : ||    ||  |   : ||    ||  ||   |  :   | |     ||  :
Db    625 PVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 679

Qy    473 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAP 531
          :|| |    : ||    || |    : ||   :||  ||  :   :   | ||
Db    680 SAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 731

Qy    532 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTL 587
             ||  |   :|||    || |   |  ::   | |   ||    ||  ||  :   :
Db    732 SSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 790

Qy    588 KEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKK 642
          | ||         :||  |   :|  :     :||  ::  :   | |     ||
Db    791 SAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESS 850

Qy    643 PAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP 697
          || :  |   : : | ||:     |:  |::|   |  || |  :  |     :| | |
Db    851 SAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYP 906

Qy    698 GVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT 757
          |  |  : ::|        |  |:     ||  |||      |     ||   :  :| |
Db    907 GSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT 966

Qy    758 --TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDR 811
            ||  || ||  ||     |   | | :   ||   : ||||||       | |||   .
Db    967 VPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA----- 1021

Qy    812 ATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPR 854
          |::||  | |                   ||   : |:|   | :|        |||
Db   1022 ---SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP- 1077

Qy    855 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHV 914
             :|:    ||: |   :  | |                |  |  |  :|   :
Db   1078 -SVTTVTNFTPTT-ITTTVCSTGT-----------NSAGETTSGCSPKTVTTTVPCST 1122

Qy    915 FMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTT 957
```

```
           | | :   |        ::    |  ||   ||   | ||
Db   1123 GTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTT 1164


RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
```

```
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23       POTENTIAL.
FT   CHAIN        24   1255       MUCIN 1.
FT   DOMAIN       24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186       POTENTIAL.
FT   DOMAIN     1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151       SEA.
FT   CARBOHYD    957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                                ISOFORM).
FT   VARSPLIC   1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                ISOFORM).
FT   VARSPLIC   1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2       T -> A (IN REF. 11).
FT   CONFLICT    134    134       P -> Q (IN REF. 9).
FT   CONFLICT    154    154       P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021       S -> T (IN REF. 3).
FT   CONFLICT   1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;
```

Query Match 12.6%; Score 651; DB 1; Length 1255;
Best Local Similarity 27.1%; Pred. No. 2.4e-19;
Matches 290; Conservative 79; Mismatches 437; Indels 266; Gaps 52;

```
Qy     75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
          | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    134 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTP----------KE 166
          |    :||:|    |    ||               | |     ||||           :
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy    167 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 206
          | | :       : ||| |   ||| :|        ||| |   ||| :|  |       |
Db    126 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 181

Qy    207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 252
          | |   ||| :|  ||   | ||| |   |  | |             :||| :   |
Db    182 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 238

Qy    253 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 289
            || |   | | |            ||| :  |        || |   |  |||
Db    239 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 298

Qy    290 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 342
            || |   | | |        || |   |  | |        |:   |: || :|  ||
Db    299 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 358

Qy    343 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 391
          | |||  |   |:| :|  |       || |   |  | |        || |   |  | |
Db    359 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 415

Qy    392 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEE 448
              || |   |  |||      |:  :| |        |       || |   |  | |
Db    416 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-- 473

Qy    449 LAPTTPEEPTPTTPEEPAPTTPKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPT 500
            |     | |   |  | | |    :||:|   |  | |        || |   |  |
Db    474 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 531

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKELAPTTTKEPTSTTSDKP 550
            |    |      :|| |   |  | |           :|||   ||       ||    :|
Db    532 AP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRP 586

Qy    551 AP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLKEPAPTTP--------- 595
            ||  | |      |:    || |   |  | |       :|| |   |  | |
Db    587 APGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP 642

Qy    596 -KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTPKEPAPTTP-------- 640
           :|||   ||     | ||     :|||  | |       :|| |   |  | |
Db    643 DTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSA 701

Qy    641 --KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPDESTPELSAEPTPKALE 691
             :||| :  | ||     | | |   | :       :  :||     |  |    :
Db    702 PDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 759

Qy    692 NSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERDLRTTPETT-------T 733
          ::|      | :  | |         |:|   :||      |:  | |  | :|       |
Db    760 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 819

Qy    734 AAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKITTLKTTTLAPKVTTTKK 791
          :||    :|       |        |:   |       | |    :|:   |  ||  |
Db    820 SAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 875

Qy    792 TITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKT 849
            |:  :    |  || |    |::  | | |   : ||      |  | | |    | :
Db    876 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHGVTSAPDTRP---APGS 930

Qy    850 TPTPRKMTSTMPELNP-----------------TSRIAEAMLQ--TTTRPNQTPNSK-- 887
          |  |    :: |:  |                   |  |  ::   |: |   || ||
Db    931 TAPPAHGVTSAPDNRPALGSTAPPVHNVTSASGSASGSASTLVHNGTSARATTTPASKST 990

Qy    888 --------------LVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDY 925
                        |   : |:: :      | :    |   |:::  : :
Db    991 PFSIPSHHSDTPTTLASHSTKTDASSTHHSSVPPLTSSNHSTSPQLSTGVSF 1042
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
```

```
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER      1      1
FT   DOMAIN       67     870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;


  Query Match           10.8%;  Score 556.5;  DB 1;  Length 875;
  Best Local Similarity  28.7%;  Pred. No. 8.9e-16;
  Matches  272;  Conservative  106;  Mismatches  375;  Indels  195;  Gaps   56;

Qy      42 TQHNKVSTSPKITTAKPINP-----------RP--SLPPNSDTSKETSLTVNKETTVET 87
           |:|  |    | : : |  |             ||  | ||    ||   | : |: :
Db       1 TKHEPVYKPVKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYG-SKTNYLPLAKKLSSYK 59

Qy      88 KETTTTNKQTSTDG--KEKTT---SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKG 142
             || | :|:     | | |   : |   |   |          |:   || ||: :
Db      60 PIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYPPTYKSKPTYKPKITYPPTYKAKPSY-- 117

Qy     143 PALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTT 193
           |:   ||:  | | |     | : ||   | | |  |: |     |:  |:: |: |:
Db     118 PSSYKPKKTYPPTYKPKLTYPPTYKPKPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSY 177

Qy     194 PKEPAPTTTKEPA--PTTPKEPA--------PTTTKEPA-PTTTKSAPTTPK--EPAPTT 240
           |    ||   :|:  ||   :|:         ||   :|  |:| |: |: |   :  ||
Db     178 P----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTY 233

Qy     241 PKKPA--PTTPKEPA--PTTPKEPT--PTTPKEP---APTTKEPAPTTPKEPA--PTAPK 289
             ||:  ||   :|:  ||   :|:  ||   :|   |  | :  ||   :|:  ||
Db     234 KAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYPPTYKA 293

Qy     290 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 339
           |:  ||   :|:  ||   :|:  ||   :|| ||   :|: | | |: ||   :|
Db     294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy     340 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 395
           |:| |: |: |    ||   :|:     :  ||   ||:  ||   :|:  ||   :|:
Db     354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy     396 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 455
             |  |  ||   :|       | |:|  |  |: |    |:   ||:     :  ||
Db     410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy     456 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 509
           :||        |:| ||     |  ||    |  | | |     |:   ||:| | |
Db     458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy     510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 565
           || |   | | | |: |    ||   | | |              || : | |    |
Db     507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy     566 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 617
           ||    |   | | |       ||   :|:  ||   ||:   |   |:| | |:   :
Db     552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy     618 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 666
           |  |: |    ||   :|:  ||   ||  |:|    |  ||    ::| | | |
Db     612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy     667 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 715
           | | :|:      |  |:|| ||   :|    :|  | |    |:       ||    |
Db     668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy     716 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 775
           | | | :  | | |  | |  | :   |   | ::|       :   |    |
Db     727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy     776 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 832
           | |: :: |     |||   |    ||: | |  :         |||  |      |  |
Db     778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826

Qy     833 TSTKKPKTMPRVRKPKTTPTP-----------RKMTSTMPELNPTSR 868
           :: |   : |   | ||:  |             :  ||  |   ||:|
Db     827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874


RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
```

```
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;
```

```
  Query Match           10.8%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity  32.6%;  Pred. No. 2.3e-15;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy      58 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 117
           |:   | ||| |   |      |   |||  |:  | : |                | | ||||
Db     445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy     118 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 172
           |     ||       |  |    |: | ::| |   || | |  :   ::||     || :
Db     488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy     173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 231
             || | |  ||      ||   |      |:|  |||| |   |  :||  ||   | ||
Db     540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy     232 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
             ::|:  | |   ||   |    : ::||    |   || |   || | | :  :|:|
Db     597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy     292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
            |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db     648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy     352 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 406
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| |   |
Db     702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy     407 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 456
           ::|  :| | | ||     :|:||| | |||      || | :    : |:|  | | |
Db     758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy     457 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 516
           :||:|  :| |              || | |  || | |       ||  :| | | ||
Db     815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy     517 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 570
             ::|  :  |   | |    :|::||  |        ::||   | | | ||  ::
Db     856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 628
           |  |:||   | |   :   ||:||       :  |||   :||    | |    |:
Db     912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy     629 PTTPKEPAPT 638
           | : | | |:
Db     967 PASCKSPRPS 976
```

```
RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
```

```
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL       1      ?
FT   CHAIN        ?    620       EXTENSIN.
FT   REPEAT      70     73        H-A-P-P.
FT   REPEAT     148    151        H-A-P-P.
FT   DOMAIN     229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT     229    235        1.
FT   REPEAT     236    242        2.
FT   DOMAIN     205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN     499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             10.7%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 1.1e-15;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy    136 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 184
          || ||: | | |   :| |::     | | :| |     |:|   : ||     | |
Db     24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy    185 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 235
          :    |:    |       |:    | |:   :|  | ||:    |       |:  :
Db     81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy    236 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 290
          |       |:|:     |      | |:   | : : |:| ::   | | |  :| ||
Db    141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy    291 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 345
          |:|    :| ||    | | |  :|:|: |     :| | | : || | :   ||   |
Db    194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy    346 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 399
             |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db    250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy    400 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 449
          | ||   :|  || : |   :|  |  |  |::       |   :|  |   :|  |  |
Db    310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy    450 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 506
          |  |  | | :   | ||  ::  |     :|  ||| |   | ||    | | |  :
Db    370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy    507 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 566
          |       | |   | |      | |    |  | |       :| |  |  : |     |
Db    428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy    567 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 626
          | :|    |       | |  | |      | |     |   : || |    |:| |
Db    482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy    627 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
            |     | |    |: | ||  : | | |      |     |   |: |   ||
Db    539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy    685 PTPKALENSPKEPGVPTTKTPAATKP 710
```

```
          |:|           ||| :| :  |
Db    599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
```

```
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;
```

Query Match 10.3%; Score 533; DB 1; Length 1087;
Best Local Similarity 27.5%; Pred. No. 8.7e-15;
Matches 216; Conservative 81; Mismatches 330; Indels 158; Gaps 38;

```
Qy     21 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 75
          :|   |     | :|      |  ||  | || :   |   :            : |||
Db    404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy     76 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 135
          :   :   :   |  |  :    |:|   : :  :  |:  |   :| ::  | |  :
Db    450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy    136 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 195
           ::  |  | :  :  :|   | ||    :  :|   || |   | |    : : | :|
Db    510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy    196 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 245
          || :|   | ||   || ||  | | |    :|:  || |   | ||     |   | ||
Db    565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy    246 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 290
             |   | ||    | | | |    :|:  | ||   :  :|   | ||     |   |
Db    620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy    291 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 342
          |    :|  || ||    | ||    :  ||   | ||     |   || |   | ||
Db    677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy    343 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 397
          : :|  :| |  ||   | ||    |   | |       : | |  || | :  :  :|
Db    730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy    398 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 457
           | |      | ||     | |  :| |     |||  |   : |   || :   | | :
Db    790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy    458 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 505
           | ||:  :|   || :|   || | |:    ||:    :|        |||     :|
Db    849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy    506 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 563
             |||  |  : |: ||    ||  :|   | |::|  :|::  |       |    :
Db    905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964
```

```
Qy    564 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 623
          : ||    || |   ||    ||  |   |  |:|:     || |: |
Db    965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP------------ 1006

Qy    624 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 683
                 ||  |  |:|                    | :| ||  : |:|
Db   1007 --------KKEEMPAAPEK-----------------KDTKEEKTTESRKPEEK------ 1034

Qy    684 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 743
             || :|   ||     :| |  :||:     |  :|::     ||      | |:: |
Db   1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy    744 TTTEK 748
          |  ||
Db   1083 TKGEK 1087
```

```
RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
```

```
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;


  Query Match           10.3%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 1.1e-14;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy     152 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 210
           ||:|| :  | :||  |  :  || :||  || ::    |:|| : :   |    |:||
Db     599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy     211 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 257
           :  |  |   |  ::  | |:|| :  | :||  |       ||:|| : :    |:||
Db     652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy     258 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 316
           : :   ||  || ::    |:|| : :       | :||  || ::      |     :
Db     712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy     317 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 372
             ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :   |:
Db     765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy     373 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 427
           ||    | :||  |  ::      |     | : |  || ::   | |:||       ||
Db     818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy     428 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 477
           :|| :  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db     877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy     478 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 537
           |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db     936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

Qy     538 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 596
             || :|  | :|     |:|| :  | :||  || ::    |  | :   | :||
Db     984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy     597 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 652
           |     | ::  |  || :|  | :||       :| :||  || ::    |  || ||
Db    1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy     653 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
           :::  |||:|  : :: | :|  ||| : ||  ||                 |||  :
Db    1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy     712 MT 713
           |
Db    1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;


  Query Match           10.3%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 9.5e-15;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy     123 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTTIKSAPTT 178
           :| |  :| | ||    |      ||   |       |  | : ||  | ||     | |
Db       8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy     179 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 233
             |||    :|  ||    ||    | |     || :    |          |||  |
Db      62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy     234 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 279
```

```
Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    280 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 331

Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    332 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 387

Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    388 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 428

Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    429 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 484

Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    485 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 533

Db    414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    534 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 585

Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    586 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 631

Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    632 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 681

Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 739

Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    740 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 792

Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    793 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 852

Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy    853 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 895

Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864
```

```
RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL       1     20       POTENTIAL.
FT   CHAIN       21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN      21     41       NONREPETIVE LINKER.
FT   DOMAIN     124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                               P-[PST]-[ST].
FT   DOMAIN     184    192       NONAPEPTIDE 1.
FT   DOMAIN     213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;
```

```
  Query Match           10.1%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity 29.2%;  Pred. No. 2.7e-14;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy     19 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 69

Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy     70 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 129

Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    130 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 188

Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    189 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 243

Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    244 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 296

Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    297 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 351

Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    352 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 401

Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    402 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 456

Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    457 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 507
```

```
Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    508 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 564
Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    565 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 620
Db    584 -------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    621 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 669
Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    670 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 725
Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    726 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 785
Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    786 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 840
Db    786 I-------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy    841 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 882
Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871
```

RESULT 12
MUC1_XENLA
ID MUC1_XENLA STANDARD; PRT; 662 AA.
AC Q05049;
DT 01-OCT-1994 (Rel. 30, Created)
DT 01-OCT-1994 (Rel. 30, Last sequence update)
DT 01-OCT-1994 (Rel. 30, Last annotation update)
DE INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS Xenopus laevis (African clawed frog).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC Xenopodinae; Xenopus.
OX NCBI_TaxID=8355;
RN [1]
RP SEQUENCE FROM N.A.
RC TISSUE=Skin;
RX MEDLINE=93077556; PubMed=1447205;
RA Hauser F., Hoffmann W.;
RT "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT polymorphism.";
RL J. Biol. Chem. 267:24620-24624(1992).
CC -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC -!- TISSUE SPECIFICITY: SKIN.
CC -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC ---

CC ---
DR EMBL; L02115; AAA74725.1; -.
DR PIR; A45155; A45155.
DR HSSP; P04002; 1WFA.
DR InterPro; IPR000519; P_trefoil.
DR Pfam; PF00088; trefoil; 6.
DR SMART; SM00018; P; 6.
DR PROSITE; PS00025; P_TREFOIL; 6.
KW Repeat; Amphibian skin; Glycoprotein; Alternative splicing.

FT	NON_TER	1	1
FT	DOMAIN	81	144	8 X 8 AA APPROXIMATE TANDEM REPEATS, ALA/THR-RICH.
FT	REPEAT	81	88	1-1.
FT	REPEAT	89	96	1-2.
FT	REPEAT	97	104	1-3.
FT	REPEAT	105	112	1-4.
FT	REPEAT	113	120	1-5.
FT	REPEAT	121	128	1-6.
FT	REPEAT	129	136	1-7.
FT	REPEAT	137	144	1-8.
FT	DOMAIN	161	202	P-TYPE 1.
FT	DOMAIN	218	301	8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT	REPEAT	218	224	2-1.
FT	REPEAT	225	239	2-2.
FT	REPEAT	240	249	2-3.
FT	REPEAT	250	259	2-4.
FT	REPEAT	260	275	2-5.
FT	REPEAT	276	287	2-6.
FT	REPEAT	288	294	2-7.
FT	REPEAT	295	301	2-8.
FT	DOMAIN	306	347	P-TYPE 2.
FT	DOMAIN	353	394	P-TYPE 3.
FT	DOMAIN	402	522	12 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT	REPEAT	402	411	3-1.
FT	REPEAT	412	419	3-2.
FT	REPEAT	420	431	3-3.
FT	REPEAT	432	443	3-4.
FT	REPEAT	444	453	3-5.
FT	REPEAT	454	460	3-6.
FT	REPEAT	461	472	3-7.
FT	REPEAT	473	479	3-8.
FT	REPEAT	480	491	3-9.
FT	REPEAT	492	498	3-10.
FT	REPEAT	499	515	3-11.
FT	REPEAT	516	522	3-12.
FT	DOMAIN	525	566	P-TYPE 4.
FT	DOMAIN	572	613	P-TYPE 5.
FT	DOMAIN	620	661	P-TYPE 6.
FT	DISULFID	162	188	BY SIMILARITY.
FT	DISULFID	172	187	BY SIMILARITY.
FT	DISULFID	182	199	BY SIMILARITY.
FT	DISULFID	307	333	BY SIMILARITY.
FT	DISULFID	317	332	BY SIMILARITY.
FT	DISULFID	327	344	BY SIMILARITY.
FT	DISULFID	354	380	BY SIMILARITY.
FT	DISULFID	364	379	BY SIMILARITY.
FT	DISULFID	374	391	BY SIMILARITY.
FT	DISULFID	526	552	BY SIMILARITY.
FT	DISULFID	536	551	BY SIMILARITY.
FT	DISULFID	546	563	BY SIMILARITY.
FT	DISULFID	573	599	BY SIMILARITY.
FT	DISULFID	583	598	BY SIMILARITY.
FT	DISULFID	593	610	BY SIMILARITY.
FT	DISULFID	621	647	BY SIMILARITY.
FT	DISULFID	631	646	BY SIMILARITY.
FT	DISULFID	641	658	BY SIMILARITY.
FT	VARIANT	276	276	K -> E.
FT	VARIANT	354	354	C -> R.
FT	VARIANT	415	415	T -> A.

SQ SEQUENCE 662 AA; 67774 MW; F085277F1ED2FD40 CRC64;

Query Match 9.9%; Score 509.5; DB 1; Length 662;
Best Local Similarity 28.1%; Pred. No. 4.9e-14;
Matches 223; Conservative 48; Mismatches 276; Indels 247; Gaps 30;

```
Qy      92 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 151
Db       3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy     152 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 211
Db      52 --------AAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy     212 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 269
Db     104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy     270 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 314
Db     164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy     315 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 374
Db     219 ---------APTTIQIATTTT---TPTTT------------TTTTKATPTT--------- 245

Qy     375 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 434
Db     246 ------------TTTTKATPTTT------------TTTKATTTTT----TPTT----- 270

Qy     435 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 494
Db     271 ---------TTTTTKATTTPTTTTTTTP------------------------------ 289

Qy     495 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 554
Db     290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy     555 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 604
Db     329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy     605 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 664
Db     385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTTPTTT 436

Qy     665 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 724
Db     437 TTTTT---------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy     725 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 784
Db     464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy     785 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 840
Db     511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy     841 MPRVRKPKTTPTPR 854
Db     567 LPQVADCKVAPSSR 580
```

```
RESULT  13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-REPEAT REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;
```

```
  Query Match             9.8%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 2e-13;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
```

```
          :||:   :||:   :||:   :||:|   |||:   :||:|   :||:|   :||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
          :||:|   :||:   :||:|   :||:|  | ||:|   :||:|   :||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
          ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
          :|   :||:|| |          ||:|   |:  |  ||: |   | || | :||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy    560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
          ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
          :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    677 STPE 680
          |  |
Db   1966 SDEE 1969
```

```
RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;
```

```
  Query Match             9.7%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 3.1e-13;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;

Qy    151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
          |:|     || |   :  ||    :|:      |     :|:   :     | :
Db   1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy    205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
          | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :||:|   :|
Db   1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy    263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
          |:|   :||:   :||:|   :||:|   :||:|   :|||:|           |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
          :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
          :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
          ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
          :|   :||:|| |          ||:|   |:  |  | : |   | || | :||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy    560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
          ||:|   :|||||  :||:|  |   |     ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
          :||:|   :||:||    :||:|   :||:|      |  :::|           |||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    677 STPE 680
          |  |
Db   1966 SDEE 1969
```

```
RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;

  Query Match             9.6%;  Score 497;  DB 1;  Length 831;
  Best Local Similarity  27.6%;  Pred. No. 1.8e-13;
  Matches  195;  Conservative   59;  Mismatches  307;  Indels  146;  Gaps   35;

Qy      37 PDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQ 96
           |  ||  | || :  ||   :             : ||| :   :   ::  |  |  :
Db     197 PSMST--HIKVKSEEKIKVVE----------KSEKETVIVEEQTEEIQVTEEVTEEED 242

Qy      97 TSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----T 152
             |:|:  : :  :    ||     | ::  | |    |  || |     | |    :
Db     243 KEAQGEEEEEAEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKS 293

Qy     153 PTTPKEPA-STTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTK 203
           |   | || : :| |     :  :|   | ||    |   || |   | ||   _| | |
Db     294 PAEAKSPAEAKSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVK 352

Qy     204 EP----APTTPKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA-- 245
           |    :|   | ||    |   | ||    | : || |   | ||    |   | ||
Db     353 SPGEAKSPAEAKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATV 411

Qy     246 --PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTA 287
           |   | ||    | | | |    :|   | || : | |    :|   | ||    |
Db     412 KSPVEAKSPAEVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVV 470

Qy     288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEP 338
           | ||    | ||    |  | :  :  ||   | ||    |   ||     :|   | |
Db     471 AKSPAEA--KSPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSP 528

Qy     339 APTTTKS-----APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTT 389
             ||     :|   | |     | ||    |   | |    |   | |     :
Db     529 VKEGAKSLAEAKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPE 588

Qy     390 PKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEEL 449
           | ||    |:||     |   :  || | :  |: | |  ||:|| |  :|   ||:
Db     589 AKPPAKEEAKRPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV 644

Qy     450 APTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTT 509
             :||     ||  | |||     : |: ||   ::|     |||    ||::
Db     645 ---KAKEPPKKVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEA 700

Qy     510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPT 569
```

```
          |  |    ||     |::  | :|      :|     |  |           |||:
Db    701 KKEEA----KEKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA----------KEPSKP 744

Qy    570 TPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAP 629
          : ||     ||       ||  |  |:|   ||           ||  |     ||  |
Db    745 SEKE----KPK-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAK 784

Qy    630 TTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
             ::       |  |  |  |    :| |: |  |:     |:|::
Db    785 AKEED-------KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824
```

Search completed: April 26, 2002, 16:19:35
Job time: 360 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:16:52 ; Search time 49.78 Seconds
(without alignments)
437.589 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1167
Perfect score: 5155
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....GKPVDGLTTLRNGTLVAFRG 968

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	10.1	805	4	US-09-103-429A-4	Sequence 4, Appli
2	506.5	9.8	744	6	5202236-25	Patent No. 5202236
3	504	9.8	1837	3	US-08-928-361B-5	Sequence 5, Appli
4	498.5	9.7	786	4	US-09-103-429A-3	Sequence 3, Appli
5	489	9.5	1721	3	US-08-700-651-5	Sequence 5, Appli
6	489	9.5	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	9.5	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	9.5	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	9.2	652	6	5202236-13	Patent No. 5202236
10	452	8.8	960	4	US-09-219-849-5	Sequence 5, Appli
11	423.5	8.2	1867	2	US-08-479-537A-5	Sequence 5, Appli
12	423.5	8.2	1867	4	US-09-083-116-5	Sequence 5, Appli
13	423.5	8.2	2035	2	US-08-479-537A-2	Sequence 2, Appli
14	423.5	8.2	2035	4	US-09-083-116-2	Sequence 2, Appli
15	419.5	8.1	2476	2	US-08-276-967-2	Sequence 2, Appli
16	419	8.1	1185	4	US-09-041-886-23	Sequence 23, Appl
17	417	8.1	829	1	US-08-642-255-132	Sequence 132, App
18	417	8.1	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	8.1	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	8.1	837	1	US-08-477-509B-103	Sequence 103, App
21	417	8.1	837	1	US-08-642-255-101	Sequence 101, App
22	417	8.1	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	8.1	837	3	US-08-482-085B-103	Sequence 103, App
24	417	8.1	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	8.0	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	8.0	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	7.8	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	7.8	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	7.8	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	7.7	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	7.7	682	1	US-08-397-633A-36	Sequence 36, Appl
32	381	7.4	1848	4	US-08-296-791-6	Sequence 6, Appli
33	381	7.4	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
34	376	7.3	960	4	US-09-219-849-6	Sequence 6, Appli
35	374.5	7.3	1537	1	US-08-325-267A-2	Sequence 2, Appli
36	371.5	7.2	1231	3	US-08-904-263A-4	Sequence 4, Appli
37	371	7.2	761	2	US-08-707-237A-84	Sequence 84, Appl
38	371	7.2	762	1	US-08-642-255-114	Sequence 114, App
39	371	7.2	762	1	US-08-397-633A-26	Sequence 26, Appl
40	370.5	7.2	1064	1	US-08-642-255-62	Sequence 62, Appl
41	368	7.1	762	1	US-08-642-255-120	Sequence 120, App
42	368	7.1	762	1	US-08-397-633A-31	Sequence 31, Appl
43	367.5	7.1	1187	1	US-08-320-559-28	Sequence 28, Appl
44	367.5	7.1	1187	3	US-08-545-860D-28	Sequence 28, Appl
45	367.5	7.1	1187	5	PCT-US94-04496-28	Sequence 28, Appl

ALIGNMENTS

```
RESULT 1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-4


  Query Match           10.1%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity  27.8%;  Pred. No. 5.2e-26;
```

```
  Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy    125 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 179
          |: |    |:     | |||  |    | | |     | |  |          |
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    180 KEPAPTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 227
          :: || |             :  | |  || ||    || | :  |||||:  |||||
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    228 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 287
           ||||  :   ||          ||||  : || |     | ||: |  ||
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 347
          ||||  : ||||  : |||||:         |||||: |||||:  ||||   |
Db    157 ---QAPTTTTQ-APTTTQ-APTTTQ---------APTTTQ-APTTTQ--APTTITQAA 198

Qy    348 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 391
          |||   :|  |  || |||    || |||  ||||:    ||  |
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    392 ------------------------------------------------EPAPTTTK 399
                                                           |||  |:
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    400 K----------------PA---------------------------------------- 402
                             ||
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    403 --------------------------------PTAP--KEPAPT-----------TPKET 417
                                           || |   || |           | |:
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    418 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 443
                                  :| |    :| |          :    |     ||
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    444 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 503
          |     |||      |:|   || | |  |||  |    ||    ||||     ||||
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    504 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
           :|||   |   |||     ||||     |  |:  | |  ||:  :   |||      ||
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    561 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 619
          ||      |||       |:|   ||||   ||| |           | |  ||:  |:  |
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    620 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 679
          ||      |  | |   ||||   ||| |  | ||| :  : |     ||  |       |
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    680 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
            :| |             | ||  || | ||
Db    686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
5202236-25
;Patent No. 5202236
;     APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;     TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;     NUMBER OF SEQUENCES: 39
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER: US/07/528,762
;       FILING DATE: 25-MAY-1990
;       APPLICATION NUMBER: 82,456
;       FILING DATE: 07-AUG-1987
;       APPLICATION NUMBER: 933,945
;       FILING DATE: 24-NOV-1986
;       APPLICATION NUMBER: 650,128
;       FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
;       LENGTH: 744
5202236-25


  Query Match             9.8%;  Score 506.5;  DB 6;  Length 744;
  Best Local Similarity   29.9%;  Pred. No. 4.6e-25;
  Matches  239;  Conservative   78;  Mismatches  308;  Indels  175;  Gaps   50;

Qy     51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
          ||:|      |:|| ||                | ::| |           | | |
Db     24 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 53

Qy    111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
                    || |  |||:          |    ||:  | | |     | :  ||
Db     54 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
            | | || ||  ||   | | |: |: |    |  |  |    ||  | | |  | |
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 275
          | |  |: |  |:  | |   || ||   | |    |  |    ||:  | | :|
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    276 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 324
            ||   :|:  ||   ||:  ||   :|:  ||   :|:  ||   :|| ||   :|:
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    325 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 377
           |:| |: |:  |    |:| |: || |    ||   :|:  ||   :|    ||   ||
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    378 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 429
          :  ||   :|:  ||   :|:  ||   ||:  ||   :|:      :  ||   | |  :
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    430 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 478
            |  |: |   |  |: |    ||   :|:  ||   :|: | | ||    |:|    |
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    479 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 532
          |     ||   | |     |  |:     | : | |   | : |:| |  :   |   |
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    533 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 591
            ||   :||  :: |  ||      |: |    |:|   :|:       ||   :|:
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    592 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 646
           ||  || |   ||   ||  ||  ||: :   :  |: |    ||   ||:  ||
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    647 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 693
              |  ||   :: | | | |      | |    |       :: |   |:|| |   :
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    694 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 739
          |  |  || |   :       ||    | | | | :   | | |  | |          | |
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719

Qy    740 KETATTTEKTTESKITATTT 759
          :   |   | ::| |  :|
Db    720 YKAKPTYPSTYKAKPTYPST 739


RESULT   3
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
```

```
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             9.8%;  Score 504;  DB 3;  Length 1837;
  Best Local Similarity   22.3%;  Pred. No. 1.8e-24;
  Matches  279;  Conservative   81;  Mismatches  481;  Indels  412;  Gaps   40;

Qy      55 TAKPINPRPSLPPNSDTSKETSLTVNKET------TVETKETTTTNKQTST-----DGK- 102
           |  | :| |: | |  |    | :  |           : || ||    :|     | |
Db     215 TGLPTDPYPNCPFNPVTGNLVSRSTGKTIPNTYAGVYRSNETKTTEPSANTNFLLVDPKI 274

Qy     103 ----EKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE 158
               | :: |  :  |   :  |  |  | |    |||     ||    | ||
Db     275 NAPCNSENSFEQVQIFDMGSKVYIPYTKCVGVKHTTTTTTT-----TTTTTTTTTTTT 329

Qy     159 PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTT 218
             :||    | ||  :  ||      ||| :  ||       |||     || |:|  |||
Db     330 TTTTTTTTTTTTTTTTTTTT------TTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTT 383

Qy     219 KEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT 278
             ||| :  ||      ||      ||      || |  | ||      ||     |
Db     384 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTT 443

Qy     279 TPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEP 338
           |       |  |||  ||      ||      |||   : ||      ||| :| ||||:|
Db     444 TTTTTTTTTTKKPTTTT------TTTTTTTTTTTTTTTTTTTTTTTTTTTTATTTTKKP 497

Qy     339 APTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKK---------------------- 376
             ||| :  || |:|: |||     || :  :   | :
Db     498 TTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECEAKGATYVGVIG 557

Qy     377 ------------------------------------------------------------ 376
```

```
Db     558 KDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPDRSLDFTIPPVA 617

Qy     377 ------------------------------PAPTTPKE-----------------PAPT 388
                                           ||  | |
Db     618 GHNSCSIIVGVSGGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCTAKYGAIHSGYQ 677

Qy     389 TPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEE 448
           |  :   ||| ||  |     ||  |  |   :| | | ||  |   ||       |
Db     678 TSADFVTTTTAKPTTTT--TGAPGQPTTTTTGSPSKPTTTTTTKATTTT------TILNP 729

Qy     449 LAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT 508
           :  || ::|| ||       ||      |  |      || :|  ||    | ||    | ||
Db     730 IITTTTQKPTTTT------TT---KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPT 780

Qy     509 TPKGTAPTTLKEPAPTTPKKPAPK----ELAPTTTKEPTS--------------TTSDK 549
           |       || |    ||   | |     |: |   ::                  ::|:
Db     781 T-----TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDE 835

Qy     550 PAPTTPKG-TAPTTPKEPA--------PTTPKEPAPTTPK------------------- 580
           | | :  |  | |:   |          |  |     |  ||
Db     836 PIPGSQAGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSY 895

Qy     581 --------------GTAPTTLK----EPAPTTPKKPAPKELAPTTT-------------- 608
                         |    ||     :|    |  |   ||    :
Db     896 LAAKNLTVDTDETYGLPIDTLTGYPLDPVSLIPFNPETGELFDPISDEIMNGTIAGIVSG 955

Qy     609 -KGPTSTTSDKPAPTTPKE-----------TAPTTPKEPAPTTPKKPAPTTPETP----- 651
                |  | | ||  |                |     |    | :  | :||
Db     956 ISASESLLSQKSAPIDPATNMVVGEFGGLLNPATGVMIPGSLGPSEQTPFSPEIEDGGII 1015

Qy     652 PPTTSEVSTPTTTKEPTTIHKSPDESTPE----------------------LSAEPTPK 688
           ||   ||:      |   :|  |  || ||                        :|  | |
Db    1016 PP---EVAAANADKFKLSIPPSVPESIPEKDQKIDSISELMYDIESGRLIGQVSKRPIPG 1072

Qy     689 ALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA----T 744
           ::          |  |||  |         |  | : :    : ||  |     |
Db    1073 SIAGDLN----PIMKTPTQT---------DSVTGKPI----DPTTGLP-FNPPTGHLINP 1114

Qy     745 TTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK--VTTTKKTITTTEIMNKP 802
           |   | :|         |   :    : :   : | | ||  |:      || |:::
Db    1115 TNNNTMDSSFAGAYKYAVSNGIKTDNVYGLPVDEITGL-PKDPVSDIPFNSTTGELVD-- 1171

Qy     803 EETAKPKDRAT----------------NSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRK 846
             | || :  |                |     | : |     :  : :   |:
Db    1172 PSTGKPINNYTAGIVSGKRGLPPIEDENGNLFDPSTKLPIDGNNQLVNPETNSTVSGSTS 1231

Qy     847 PKTTPTPRKMTS---TMPE--------------LNPTSRIAEAMLQTTTRPNQTPNSKL 888
             | | |    :    :|:                : ||: | :  :  |   | | |
Db    1232 GSTKPKPGIPVNGGGVVPDEEAKDQADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNI-- 1289

Qy     889 VEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPD--------MDYLPRVPNQGIIINP-- 938
             :|| |       |  | | |    |: | :          :  ||  |: | ||:|
Db    1290 --INP--ETGKVIPGSLPGSLNYPSFNTPQQTDEITGKPVDTVTGLPYDPSTGEIIDPAT 1345

Qy     939 ------------MLSDETNICNGK---------------PVDGLTTLRNGTLV 964
                       :|::  ||   :                || ||  | ||||
Db    1346 KLPIPGSVAGDEILTEVLNITTDEVTGLPIDLETGLPRDPVSGLPQLPNGTLV 1398
```

```
RESULT   4
US-09-103-429A-3
; Sequence 3, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
```

```
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  3:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  786 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             9.7%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 1.6e-24;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 191
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    192 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 251
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    252 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 308
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    309 ----------------------------------------------PTTTKEPSPTTPK- 321
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    322 ----EPAPTTT---------------------------------------------- 328
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    329 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 361
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    362 APTT-----PKEP-----------------------APTTPKKPAPTT----------- 381
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    382 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 438
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    439 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 496
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 556
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    557 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 616
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    617 DKPAPTTPKETAPTTPKEPAPTTPKKP 643
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   5
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5


  Query Match             9.5%;  Score 489;  DB 3;  Length 1721;
  Best Local Similarity   22.1%;  Pred. No. 1.5e-23;
  Matches  284;  Conservative   82;  Mismatches  476;  Indels  442;  Gaps   41;

Qy      5 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 64
Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy     65 LPPNSDTSKE--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTT---------- 208

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
Db    209 ---------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    170 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 229
Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    230 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 289
Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTT---------- 350

Qy    290 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 349
Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTKKPTTTTTATTTTT 406

Qy    350 PKE-------------------------------------------PSPTT------- 357
Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    358 ---------------------------TKEP--------------------APTTPKE 368
```

```
Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    369 PAPTTPKKPAPTTPKE----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 410

Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    411 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 470

Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    471 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 530

Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPL 680

Qy    531 PK----ELAPTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 570

Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    571 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 627

Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    628 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 673

Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    674 ------------------------PDEST------------PELSAEPTPKALENSPK 695

Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    696 EPGVPTTKTPAATKPEMTTTAKDKTTER----------------DLRTTPETTTAAPKM 738

Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    739 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797

Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    798 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 842

Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    843 ---------------RVRKPKTTPTPRKMTS--TMPE---------------------- 862

Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    863 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 917

Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    918 EVTPD--------MDYLPRVPNQGIIINP--------------MLSDETNICNGK----- 950

Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy    951 ----------PVDGLTTLRNGTLV 964

Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281
```

```
RESULT   6
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6
```

Query Match 9.5%; Score 489; DB 3; Length 1721;
Best Local Similarity 22.1%; Pred. No. 1.5e-23;
Matches 284; Conservative 82; Mismatches 476; Indels 442; Gaps 41;

```
Qy      5 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 64

Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy     65 LPPNSDTSKE--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109

Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTT---------- 208

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169

Db    209 --------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    170 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 229

Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    230 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 289

Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT---------- 350

Qy    290 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 349

Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTKKPTTTTTATTTTT 406

Qy    350 PKE----------------------------------------------PSPTT------- 357

Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    358 ----------------------------TKEP--------------------APTTPKE 368

Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    369 PAPTTPKKPAPTTPKE----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 410

Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580
```

```
Qy    411 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 470
Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    471 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 530
Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPL 680

Qy    531 PK----ELAPTTTKEPTS----------TTSDKPAPTTPKG-TAPTTPKEPAPTT 570
Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    571 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 627
Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    628 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 673
Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    674 ----------PDEST----------PELSAEPTPKALENSPK 695
Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    696 EPGVPTTKTPAATKPEMTTTAKDKTTER----------DLRTTPETTTAAPKM 738
Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    739 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    798 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 842
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    843 --------------RVRKPKTTPTPRKMTS--TMPE---------------------- 862
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    863 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 917
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    918 EVTPD--------MDYLPRVPNQGIIINP----------MLSDETNICNGK----- 950
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy    951 ----------PVDGLTTLRNGTLV 964
Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281
```

RESULT 7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
; GENERAL INFORMATION:
; APPLICANT: Hoffman, Stephen L.
; APPLICANT: Charoenvit, Yupin
; APPLICANT: Hedstrom, Richard
; APPLICANT: Khusmith, Srisin
; APPLICANT: Rogers IV, William O.
; TITLE OF INVENTION: Protective malaria sporozoite surface protein
; TITLE OF INVENTION: immunogen and gene
; NUMBER OF SEQUENCES: 2
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: A. David Spevack
; STREET: NMRDC Building 1 T-12 National Naval
; STREET: Medical Center
; CITY: Bethesda
; STATE: MD
; COUNTRY: USA
; ZIP: 20814-5044
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.24
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/638,431
; FILING DATE: 19910110
; CLASSIFICATION: 424
; ATTORNEY/AGENT INFORMATION:
; NAME: Spevack, Avrom D.
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (301) 295-6759
; TELEFAX: (301) 295-4033
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 826 amino acids
; TYPE: AMINO ACID
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-07-638-431-2

Query Match 9.5%; Score 488.5; DB 1; Length 826;
Best Local Similarity 26.4%; Pred. No. 7.2e-24;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
Db    607 PKEPI-------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    685 PTPKALENSPK-------------------EPGVPTTK 703
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745
```

RESULT 8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018

```
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene encoding
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  PCT/US92/00018
;      FILING DATE:  19920103
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avram D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
PCT-US92-00018-2


  Query Match             9.5%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 7.2e-24;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy     168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
           ||  ::  |  |  |  |       |: |:||  |    :|     |  |  |  |
Db     273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy     221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
           |      : |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  :|:     |
Db     333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy     277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
             ||   |  ||  || |  |    |: | :| |       |: |  :| |    : |:
Db     393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy     335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
             :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db     444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy     393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
           |:       |  |  |  |: | |  |: |||              |: | |  |
Db     493 PS-------NPNEPSNPNEPSNPNE--PSNPKK----------------PSNPNE--P 523

Qy     452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
           : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db     524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy     511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
                     |: |::| |:|  |:  |||    |:   |  |:   |  |  |  |  :
Db     565 ------------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
           ||||               |  : ||:|    : |   : |     :  |       |
Db     607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy     631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
           |  |           |: |: :  | | |  :  ::  | |    |   |:|  |
Db     650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy     685 PTPKALENSPK-------------------EPGVPTTK 703
           |     |  ||                   || :|  :
Db     707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   9
5202236-13
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;      LENGTH: 652
5202236-13


  Query Match             9.2%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 3.3e-23;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy      51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
           ||:|      |:|| ||                | ::| |           | | |
Db      11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy     111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
                     || |  |||:          |    ||:  | | |     | :  ||
Db      41 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy     167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
             | | || ||  ||   | | |: |: |    |  |  |    ||  | | |  |  |
Db      80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy     224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 278
             | |  |: |  |:  | |   ||  ||  | |    | |    | :  | | :|  |
Db     139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196

Qy     279 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 329
             |    ||   ||:  ||   :|:  ||   :|:   | :  |: |     |   | | |
Db     197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy     330 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 376
           : ||   :|         |:| |: || |    ||   :|:  ||   :|    ||  |
Db     253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308

Qy     377 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 432
           |:  ||   :|:  ||   :|:   | |  |: |    ||   : :      |: ||   |
Db     309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy     433 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 487
             || |   |  |: |    ||   :|:  ||   :|: | | | | |: |    ||
Db     363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy     488 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 540
           :|  |:|    |  |    || : | |    | : |:| |  :   |   |    ||
Db     414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473
```

```
Qy    541 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 598
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    599 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 651
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    652 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    712 MTTTAKDKTTERDLRTTPETTTAAP 736
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645


RESULT 10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
;  APPLICANT:  VAN HEERDE, GEORGE V.
;  APPLICANT:  VAN RIJN, ALEXIS C.
;  APPLICANT:  BOUWSTRA, JAN B.
;  APPLICANT:  DE WOLF, FREDERIK A.
;  APPLICANT:  MOOBROEK, ANDREAS
;  APPLICANT:  WERTEN, MARC W.T.
;  APPLICANT:  WIND, RICHELE D.
;  APPLICANT:  VAN DEN BOSCH, TANJA J.
;  TITLE OF INVENTION:  SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
;  TITLE OF INVENTION:  SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
;  TITLE OF INVENTION:  PREPARATION THEREOF
;  FILE REFERENCE:  2728-2
;  CURRENT APPLICATION NUMBER: US/09/219,849
;  CURRENT FILING DATE:  1998-12-23
;  NUMBER OF SEQ ID NOS:  50
;  SOFTWARE:  PatentIn Ver. 2.1
; SEQ ID NO 5
;   LENGTH: 960
;   TYPE: PRT
;   ORGANISM: Artificial Sequence
;   FEATURE:
;   OTHER INFORMATION: Description of Artificial Sequence: Illustrative
;   OTHER INFORMATION: amino acid sequence
US-09-219-849-5


  Query Match             8.8%;  Score 452;  DB 4;  Length 960;
  Best Local Similarity   25.6%;  Pred. No. 1.8e-21;
  Matches  173;  Conservative   60;  Mismatches  309;  Indels  134;  Gaps   41;

Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 189
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    190 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 245
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    246 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA---------------- 284
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    285 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 330
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    331 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 387
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    388 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 437
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    438 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 492
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    493 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 535
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    536 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 590
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    591 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 642
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    643 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 693
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    694 PKEPGVPTTKTPAATK 709
Db    696 SRDPGPPGAHGPAGPK 711


RESULT 11
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
; GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;
```

```
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             8.2%;  Score 423.5;  DB 2;  Length 1867;
  Best Local Similarity   24.4%;  Pred. No. 2.4e-19;
  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||               | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
             :||| :|        |||     |             ||     |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|        |||     | | |         ||     |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |        ||     | |||       ||     |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||      |||     |  |         |||     |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
           ||      |  | |       ||      |  | |       ||      |  |||     |

Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  : |    | |     |:    ||     |  | |     |: : |     |  |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
             | |    :||:    |  | |     |:  :  |  |  |||     | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||      ||    :| |  | |     |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             | | |       :||      |  | |           :| |   ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||      |  |     : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||     |:    :           |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                 |       |   |  ||      :  :     |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
             : ||      |  |     |      | |    :||  |:  |      |      |: |:
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy     883 TP 884
             |
Db     925 RP 926


RESULT  12
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
```

```
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5

  Query Match             8.2%;  Score 423.5;  DB 4;  Length 1867;
  Best Local Similarity   24.4%;  Pred. No. 2.4e-19;
  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |   ||                | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
               :||| :|        |||     |            ||     |  | |      |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
           :  | | :|         |||      |  | |        ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
           :  | | :  |         ||     |  |||        ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
           :  |  |   ||       |||     |  |          |||     |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
           ||      |  | |        ||      |  | |        ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |     | |      |:     ||       |  | |      |:  :  |     | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |     :||:      |  | |      |:  :  |  |  |||      | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |    ||          ||      :| |  | |     |:     ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             | | |       :||      |  | |            :| |   ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |            :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||      |  |     : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:      :                |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                 |       |   | ||       :  :     |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
           : ||     |  |    |       | |    :||  |:  |      |      |: |:
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy     883 TP 884
           |
Db     925 RP 926


RESULT  13
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
```

```
;       STATE:  Virginia
;       COUNTRY:  United States
;       ZIP:  22313-1404
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.0, Version #1.30
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  US/08/479,537A
;       FILING DATE:  07-JUN-1995
;       CLASSIFICATION:  514
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  FR 90/13101
;       FILING DATE:  23-OCT-1990
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  2:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  2035 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1899
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;       OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2

  Query Match             8.2%;  Score 423.5;  DB 2;  Length 2035;
  Best Local Similarity  24.4%;  Pred. No. 2.6e-19;

  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |    ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
               :||| :|         |||      |                ||      |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|         |||      |  | |        ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |         ||      |  |||              ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  | |  ||        |||      | |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
           ||      |  | |        ||      |  | |        ||      |  |||      |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |      |:    ||      |  | |      |:  :  |      | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |     :||:     |  | |      |:  :  |  |  |||      | ||      |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||       ||     :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
           |  | |       :||      |  | |            :| |   ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |            :| | :  | |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
               | :         :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||     |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
               |       |   | ||      :  :    |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
             : ||      |  |   |       | |    :||  |:  |      |      |: |:
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy     883 TP 884
           |
Db     925 RP 926


RESULT  14
US-09-083-116-2
```

```
; Sequence 2, Application US/09083116
; Patent No. 6203795
; GENERAL INFORMATION:
;   APPLICANT:  CHAMBON, Pierre
;   APPLICANT:  KIENY, Marie-Paule
;   APPLICANT:  LATHE, Richard
;   APPLICANT:  HAREUVENI, Mara
;   TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;   TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;   NUMBER OF SEQUENCES:  5
;   CORRESPONDENCE ADDRESS:
;     ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;     STREET:  P.O. Box 1404
;     CITY:  Alexandria
;     STATE:  Virginia
;     COUNTRY:  United States
;     ZIP:  22313-1404
;   COMPUTER READABLE FORM:
;     MEDIUM TYPE:  Floppy disk
;     COMPUTER:  IBM PC compatible
;     OPERATING SYSTEM:  PC-DOS/MS-DOS
;     SOFTWARE:  PatentIn Release #1.0, Version #1.30
;   CURRENT APPLICATION DATA:
;     APPLICATION NUMBER:  US/09/083,116
;     FILING DATE:
;     CLASSIFICATION:
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  08/479,537
;     FILING DATE:
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  WO PCT/FR91/00835
;     FILING DATE:  23-OCT-1991
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  US 08/039,320
;     FILING DATE:  04-APR-1993
;   PRIOR APPLICATION DATA:
;     APPLICATION NUMBER:  US 08/403,576
;     FILING DATE:  14-MAR-1995
;   ATTORNEY/AGENT INFORMATION:
;     NAME:  Teskin, Robin L.
;     REGISTRATION NUMBER:  35,030
;     REFERENCE/DOCKET NUMBER:  017753-025
;   TELECOMMUNICATION INFORMATION:
;     TELEPHONE:  (703) 836-6620
;     TELEFAX:  (703) 836-2021
; INFORMATION FOR SEQ ID NO:  2:
;   SEQUENCE CHARACTERISTICS:
;     LENGTH:  2035 amino acids
;     TYPE:  amino acid
;     STRANDEDNESS:  single
;     TOPOLOGY:  linear
;   MOLECULE TYPE:  peptide
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  128..1899
;     OTHER INFORMATION:  /note= "The amino acids spanning
;     OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;     OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;     OTHER INFORMATION:  repeats varies from 1 to 40."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  134
;     OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;     OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;     OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  144
;     OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;     OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;     OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  147
;     OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;     OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;     OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;   FEATURE:
;     NAME/KEY:  Peptide
;     LOCATION:  1..21
;     OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;     OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2


  Query Match             8.2%;  Score 423.5;  DB 4;  Length 2035;
  Best Local Similarity   24.4%;  Pred. No. 2.6e-19;
  Matches  235;  Conservative   63;  Mismatches  461;  Indels  203;  Gaps   48;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
             :||| :|         |||      |                ||     |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|         |||      |  | |         ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |         ||      |  |||       ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||      |  | |         ||      |  | |         ||      |  |||      |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           : : |     | |      |:    ||      | | |      |:  :  |      | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
            | |     :||:     |  | |      |:  :  |  |  |||      | ||      |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||       ||     :| |  | |      |:     ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             |  | |       :||      |  | |           :|  |   ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
                | :       :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
             :|    :||       |:  | |  | :|  ||     |:     :            |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
```

```
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT-SRIAEAMLQTTTRPNQ 882
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPXPGSTAPXAHGVTSAPDX 924

Qy    883 TP 884
Db    925 RP 926
```

RESULT 15
US-08-276-967-2
; Sequence 2, Application US/08276967
; Patent No. 5851817
; GENERAL INFORMATION:
; APPLICANT: Hardy, Daniel M.
; APPLICANT: Garbers, David L.
; TITLE OF INVENTION: Species-Specific Egg-Binding Proteins of
; TITLE OF INVENTION: Sperm
; NUMBER OF SEQUENCES: 8
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Arnold, White & Durkee
; STREET: P. O. Box 4433
; CITY: Houston
; STATE: Texas
; COUNTRY: USA
; ZIP: 77210-4433
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.25
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/276,967
; FILING DATE: Submitted Herewith
; CLASSIFICATION: 530
; ATTORNEY/AGENT INFORMATION:
; NAME: Kitchell, Barbara S.
; REGISTRATION NUMBER: 33,928
; REFERENCE/DOCKET NUMBER: UTSD:418\KIT
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 713-787-1400
; TELEFAX: 713-789-2679
; TELEX: 79-0924
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 2476 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-276-967-2

Query Match 8.1%; Score 419.5; DB 2; Length 2476;
Best Local Similarity 32.5%; Pred. No. 5.7e-19;
Matches 163; Conservative 43; Mismatches 177; Indels 119; Gaps 31;

```
Qy    406 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 465
Db    313 PSETSVSTEKPVAPTE----KPTVPSEIYTIPTEKPMVHMEKPIVHT--EKPTVPT-EKP 365

Qy    466 APTTPKAAAPNTPKEPA----PTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 521
Db    366 TIPTEKSTVPT--KKPTVFKEPTLPPE-GPTVPAE-RPTTPPEGPAVPPKG--PTVLTE- 418

Qy    522 APTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPT-TPKEPAPTTPKEPAPTTPK 580
Db    419 ------------WPTSHTEKSTVHTEKPILPTGKSTIPTEKPMVPTKRT------TTP- 458

Qy    581 GTAPTTLKEPAPTTPKKPAPKELAPT-TTKGPTSTT---SDKPAPTTPKETAPTTPKEPA 636
Db    459 -TERTTIPAEKPTV---PIEKPMVPTERTTIPTERTTIPTEKPTVPTEKLTVPT--EKPI 512

Qy    637 PTTPKKPAPTTPETPPP-----TTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPT 686
Db    513 VPTEKPIVPTEKHTIPTEKLTVLTERTTTPTERTTIPTEKPTVPTEKPSVPT-EKPTVPT 571

Qy    687 PKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTT 746
Db    572 --------EEPTIPTEKLTVPT--ERTTTPTKRTTTPTIRTTTPTI--------RTTTPT 613

Qy    747 EKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETA 806
Db    614 ERTT------TPTIRTTTPTERTT---IPTKKTT------VPTEKTIIPT------ERTI 652

Qy    807 KPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT 866
Db    653 AP---------TTPQP-SPTLVPTQPAAVVMPST------SATTVTPRTTIASCP---PN 693

Qy    867 SRIAEAMLQTTTRPNQTPNSKL 888
Db    694 AHFERCACPVSCQ-SPTPNCEL 714
```

Search completed: April 26, 2002, 16:17:05
Job time: 216 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:19:22 ; Search time 82.85 Seconds
(without alignments)
931.380 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1212
Perfect score: 5416
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....SPIDTVFTRCNCEGKTFFFK 1013

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1062.5	19.6	3020	2	A43932	mucin 2 precursor,
2	942	17.4	1664	2	T18262	S-layer protein -
3	853	15.7	1489	2	T31108	cyst germination s
4	802	14.8	1274	2	T16251	hypothetical prote
5	789.5	14.6	1367	1	S48478	glucan 1,4-alpha-g
6	784	14.5	1188	2	S49915	extensin-like prot
7	776	14.3	2187	2	T30826	nascent polypeptid
8	672	12.4	1344	1	A35175	mucin 1 precursor,
9	671	12.4	1151	2	T18535	high molecular mas
10	659.5	12.2	1229	2	T25697	hypothetical prote
11	633	11.7	489	2	T11622	extensin class 1 p
12	632.5	11.7	7962	2	I38346	elastic titin - hu
13	632	11.7	761	2	C84672	hypothetical prote
14	629	11.6	3507	2	T34513	hypothetical prote
15	626.5	11.6	990	2	I51618	nucleolar phosphop
16	625	11.5	6642	2	T29757	protein UNC-89 - C
17	622.5	11.5	.839	2	F75518	hypothetical prote
18	622	11.5	3570	2	T45025	mucin MUC5B, trach
19	620.5	11.5	971	2	T19431	hypothetical prote
20	607.5	11.2	801	2	T29018	hypothetical prote
21	607.5	11.2	924	2	S27923	gene LF3 protein -
22	605	11.2	379	2	S50125	larval glue protei
23	559.5	10.3	350	2	S22456	hydroxyproline-ric
24	556.5	10.3	856	2	T16543	hypothetical prote
25	556.5	10.3	875	2	S23760	polyphenolic adhes
26	555.5	10.3	1630	2	A53577	ascites sialoglyco
27	553	10.2	2232	2	T34434	hypothetical prote
28	551	10.2	620	2	S06733	hydroxyproline-ric
29	545	10.1	873	2	A47283	calphotin - fruit
30	542	10.0	369	2	S20500	hydroxyproline-ric
31	538.5	9.9	416	2	JU0465	extensin precursor
32	533	9.8	1087	1	QFMSH	neurofilament trip
33	532	9.8	756	2	T27642	hypothetical prote
34	530.5	9.8	1162	2	JH0557	exo-alpha-sialidas
35	530	9.8	865	2	A47282	calcium-binding pr
36	522	9.6	328	2	JQ0985	hydroxyproline-ric
37	522	9.6	1459	2	T32271	hypothetical prote
38	518	9.6	813	2	S70795	vsaA protein precu
39	518	9.6	866	2	T45462	membrane glycoprot
40	518	9.6	1072	1	A37221	neurofilament trip
41	512.5	9.5	867	2	T45463	membrane glycoprot
42	509.5	9.4	662	2	A45155	mucin FIM-C.1 - Af
43	508	9.4	1832	2	T31113	mucin-like glycopr
44	505	9.3	700	2	A54641	interspersed repea
45	504.5	9.3	606	2	A43427	neurofilament trip

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 19.6%; Score 1062.5; DB 2; Length 3020;
Best Local Similarity 29.2%; Pred. No. 1.5e-38;
Matches 342; Conservative 74; Mismatches 442; Indels 315; Gaps 39;

```
Qy     32 FKVTTPDTSTTQH---NKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
          ::: |: : :|     ::| |   |       |: |        || |    ||  |
Db   1240 WEICGPNGTVEKHFNICSITTRPSTLTTFTTITLPTTP--------TSFTTTTTTTTPTS 1291

Qy     89 ET--TTTNK--------------QTSTDGKEKTTS--AKETQSIEKTSAKD-------- 121
           |  :|| |              | ||  :       : ||  | ||
Db   1292 STVLSTTPKLCCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQH 1351

Qy    122 ------------------------------------------LAPTSKVLAKPTPKAE 137
                                                     :   |     |:|
Db   1352 GQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPT 1411

Qy    138 TTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-TTPKEPAPTTTKSAPTTPKE 196
          |||  |  |||  || ||   | :|||  |  ||    | |||  |  |||   |||
Db   1412 TTTTLPPTTTPSPPTTTTTTPPPTTTPSPPITTTTTPLPTTTPSPPISTTTTPPPTTTPS 1471

Qy    197 PAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP-TTPKEPAPTTPKKPAPTTPKEPAPT 255
          | ||||  |  |||  |  |||  | ||||  | | |||  |  :|   |  |||  |  |
Db   1472 P-PTTTPSPPTTTPSPPT-TTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTT 1529

Qy    256 TPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEP 315
          |   |  |||  |  |   | ||:     ||    | |||   | |||   |  |||  |
Db   1530 TTTPPPTTTPSPPTTTPITP-PTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSP 1588

Qy    316 SPTTPKEPAPTTTKSAP-TTTKEPAPTTTKSAP-TTPKEPSPTTTKEPAPTTPKEPAPTT 373
             |   | |||| | | |||  | |||| | | |||  |  :||  |  |||  |  ||
Db   1589 PTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTT 1648

Qy    374 PKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKL 433
             |  |||  |  |||  |  |||  |  | |  |  |||  : |||    | |||
Db   1649 TTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTP--SPPTT----TMTTPS-- 1700

Qy    434 APTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT 493
            |||  |  ||    : |||  | |||   |:|||  :    |     |||   |  ||
Db   1701 PTTTPSSPITTTTTPSSTTTP-SPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTT 1759

Qy    494 PKEPAPTTP-----KETAPTTP------------KG------------------------ 512
          |  |: | |       | ||||             |
Db   1760 PLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAA 1819
```

```
Qy    513 ---------------------------------------------------------- 512

Db   1820 NISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECV 1879

Qy    513 TAPTTLKEPAPTTPKKPAPKELAPTT--TKEPTSTTSDKP----APT-TPKGTAPTTPKE 565
          | |||:     ||  : | |   :  ||  |  || |::  |    ||| ||  |  |
Db   1880 TQPTTM---TTTTTENPTPTPITTTTTVTPTPTPTSTQSPNGLQAPTPTPISTTTTVTPT 1936

Qy    566 PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKPAPTTP 624
          | ||  : |  |||  |  |    | ||  : |    :  |||  || | || |    ||
Db   1937 PTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTVLITTTTTMTPTPTPTSTKSTTVTP 1995

Qy    625 KETAPTTPKEPAPTTPKKP-------APTTPETPPPTTSEVSTPTTTK-----EPTTIHK 672
            |  |    | ||  : |          |   || |||     || |      | || :
Db   1996 ITTTTTVTATPTPTGTQTPTMIPISTTTTVTPTPTPTTGSTGPPTHTSTAPIAELTTSNP 2055

Qy    673 SPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETT 732
          |: |||: |   :    |::     :|      :| |  | ||   |:     : | :|
Db   2056 PPESSTPQTSRSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPTTTSGGHTLSPPPST 2115

Qy    733 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITT---LKTTTLAPKVTTT 789
          | :|  |    |||   : |  | :| | |:|:    ||  ::|   ::|| | |  ::
Db   2116 TTSPPGTPTRGTTT--GSSSAPTPSTVQTTTTSAWTPTPTPLSTPSIIRTTGLRPYPSSV 2173

Qy    790 KKTITTTEIMNKPEE----------------------------TAKPKDRATNSKATTP 820
                :   | |                              : |    | || :||
Db   2174 LICCVLNDTYYAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTPSPTPTPSK-STP 2232

Qy    821 KPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP 880
          | ||:  | |||   ||   |    |:             : | |  : :  :  | :
Db   2233 TPSKPSSTPSKPTPGTKPPECPDFDPPR------------QENETWWLCDCFM-ATCKY 2278

Qy    881 NQTPNSKLVEVNP-------------KSEDAGG------------AEGETPHMLL----- 910
          | |     ||  |              : ||  |              |: || :
Db   2279 NNTVEIVKVECEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGD-PHYVTFDGLY 2337

Qy    911 ------RPHVFMPEVTPDMDYLPRVPNQGIIIN 937
                :| : |::| :|      | |: |:
Db   2338 YSYQGNCTYVLVEEISPSVD------NFGVYID 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

Query Match 17.4%; Score 942; DB 2; Length 1664;
Best Local Similarity 33.0%; Pred. No. 1.3e-33;
Matches 276; Conservative 83; Mismatches 312; Indels 166; Gaps 43;

```
Qy    128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
          |: :| | || :    |  |   ||||:          ||||:    ||   || |:
Db    758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy    187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
            :  ||:|| || |    ||  ||     || : || |:   :   ||:|| ||
Db    805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy    243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
            ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
```

```
Db    865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy    302 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPTTTKSAPTTPKEPSPTT 357
Db    920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS----------DEPTP 968

Qy    358 TKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPAPTTPKE 416
Db    969 SDEPTPS--DEPTPS----DEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSD--- 1019

Qy    417 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAP 475
Db   1020 ---------EPTPSDE--PTPSDEPTPSD----EPTPSDEPTPSETPEEPIPTDTPSDEP 1064

Qy    476 NTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK 532
Db   1065 TPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS 1122

Qy    533 ELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA 591
Db   1123 D-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS----DEPTPSDEPT 1175

Qy    592 PT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKEPAPT-TPKKPAPT 646
Db   1176 PSETPEEP-----IPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETPEEPIPT 1229

Qy    647 TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPA 706
Db   1230 DTPSDEPTPSD--EPTPSDEPT----PSDEPTP--SDEPTP---SETPEEP-IPTDTPSD 1277

Qy    707 ATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
Db   1278 EPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPS 1337

Qy    767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ-KP 825
Db   1338 DEPTPSDEPT-----------------------PSDEPTPSDEPTPSDEPTPSDEPTP 1372

Qy    826 TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTS--------------TMPELNPTSRIA 870
Db   1373 SETPEEPTPTTTP-----TPTPSTTPTSGSGGSGGSGGGGGGGGGTVPTSPTPTPTS--- 1424

Qy    871 EAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLP 927
Db   1425 --------KPTSTPAPTEIE-EPTPSDVPGAIGGEHRAYLRGY-------PDGSFRP 1465


RESULT   3
T31108
cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1
C;Genetics:
A;Gene: car90


  Query Match             15.7%;  Score 853;  DB 2;  Length 1489;
  Best Local Similarity   32.3%;  Pred. No. 7.6e-30;
  Matches  355;  Conservative   46;  Mismatches  475;  Indels  222;  Gaps   49;

Qy      9 TKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTT------QHNKVSTSPKITTAKPINPR 62
Db    338 TPYAPTEKPYDVEETTYVTEESTYAPTKSETNAPTERMHYAHIEKPCDTEVTMYAPTEET 397

Qy     63 PSLP--------------PNSDT----SKETSLTVNKETTVETKETTT---TNKQTSTD 100
Db    398 TYAPTEETTYAPTEETTYAPTEETPYEPTEETTYTPTEETTYAPTEETTYAPTEKTTYAP 457

Qy    101 GKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPA 160
Db    458 TEETTYAPTEETPYEPTEETTYAPTKETTYAPT---EETTYASTEETTYAPTEETTYAPA 514

Qy    161 STTPKEPTPTTIKSAPTTPKEPAPT-TTKSAPTTPKEPAPT--TTKEPAPTTPKEPAPTT 217
Db    515 EETPYEPTEET-TYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEET 573

Qy    218 TKEPAPTTTKSAPT----------TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKE 267
Db    574 TYAPTEETT-YAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYA 632

Qy    268 P--------------AP---TTKEPAPTTPKEP------APTAPKKPAPTTPKEPAPTTP 304
Db    633 PTEETTYASTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEE 692

Qy    305 KEPAPT--TTKEPSPTTPKEP------APT-TTKSAPT----------TTKEP------A 339
Db    693 TTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTYA 752

Qy    340 PT-TTKSAPT--TPKEPSPTTTKEP------APTTPKEPAPTTPKKPAPTTPKEPAPT-- 388
Db    753 PTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEE 812

Qy    389 TPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEE 448
Db    813 TPYEPTEETTYAPTEETPYEPTEET--TYTPTEETTYAPTEETTYAPTEKTTYAPTEETT 870

Qy    449 LAPT--TPEEPTPTTPEEP------APT----------TPKAAAPNTPKEPAPTTPKEP- 489
Db    871 YAPTEETPYEPTEETTYAPTKETTYAPTEETTYASTEETTYAPTEETTYAPAEETPYEPT 930

Qy    490 -----APTTPKEPAPTTPKETAPTTPKGTAPT--TLKEPAPTTPKKPAPK-ELAPT--TT 539
Db    931 EETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETT 990

Qy    540 KEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEP------APTTPKGTAPT-------- 585
Db    991 YAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAST 1050

Qy    586 ----------TLKEPAPTTPKKP------APKE---LAPT--TTKGPTSTTSDKPAPTTP 624
Db   1051 EETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETT 1110

Qy    625 KETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT--TIHKSPDESTPELS 682
Db   1111 YAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTY-APTEATTYAP 1169

Qy    683 AEPTPKA-LENSPKEPG-----VPTTKTPAATKPEMTTTAKDKTTERDLRTTP----ETT 732
Db   1170 TEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPTEET 1229

Qy    733 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTT----------PFKITT---LKT 779
Db   1230 TYAP-----TEETTYEPTEETTYAPTEETTYAPTEETTYAPTEETMYAPIDETTYGPTEE 1284

Qy    780 TTLAPKVTTTKKTITTTEIMNKPEETAKPKDRAT--NSKATTPKP-QKPTKAPKKPT--- 833
Db   1285 TTYAPTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPMEETPYE 1344

Qy    834 ---------STKKP-KTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP--N 881
Db   1345 PAEESTSTVSTEKPCNTEEFTDEPTDEPTDE--PSDEPTDEPTDEPTDLPTDEPSTPCDN 1402

Qy    882 QTPNSKLVEVNPKSEDAG 899
Db   1403 QGINGIGVENKVRYNNAG 1420
```

```
RESULT   4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN00
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

  Query Match             14.8%;  Score 802;  DB 2;  Length 1274;
  Best Local Similarity   28.0%;  Pred. No. 1e-27;
  Matches  290;  Conservative  102;  Mismatches  414;  Indels  230;  Gaps   56;

Qy       6 KNRTKK-KP-----TP-----KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTS 50
           || ||| ||      ||      :||| ::    |   |     |   | |      || |
Db     274 KNPTKKWKPPWEDETPVEEVKEPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPS 333

Qy      51 PKITTAKPI-NPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
             :    |: ||      | :   | :  | | :  | |      |:         ::
Db     334 SPVVPPTPVKNPVKKYKPPWEVDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPS 393

Qy     110 ETQSIEKTSAKDLAPTSKVLAKPTPKAE----TTTKGPA-----------------LTTP 148
           ::      :| :|      :  |||:    |  | |                     :  |
Db     394 PKKAAPAVKPRDSSPKKATPLQADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQP 453

Qy     149 KEPTPTTP----KEPASTTPKEP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAP 199
           : |   ||    ||||:    :|    || |  |  |  |:: :|    : | : : ||
Db     454 EAPAKKTPVLKRKEPAAKDTAKPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA- 512

Qy     200 TTTKEPA------------------PTTPKEPAPTTTKEPAPT-------TTKSAPTTP 233
           |  | |                    ||  |:      |:|||             ||  |
Db     513 TPVKNPVKKWRPPWEDDETPADDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKP 572

Qy     234 KEP-----APTTPKKPAPTTPKE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EP 283
           ::|     ||| |:|  |   |:    || :  |||  : |::|:|    ||   || |
Db     573 RDPSPKKVAPTAPEKKTPVLAKKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEV 632

Qy     284 APTAPKKPAP------TTPKEPAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPT 333
           || | ||| |       | ||:  | :|  | ||  | |      |    : ||    | |
Db     633 APAAVKKPEPISKPKDTAPKKAEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPE 691

Qy     334 TTKEPAPTTTKSAPTTP-KEPSPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPT 388
             |:  |   | ||| |  | :      |:   ||  : ||  |: |:|    | :|||
Db     692 PEKK-TPVLAKKAPTKPDSEAAADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPK 750

Qy     389 TPKEPAPTTTKKPAPTA-PKEPAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPA 442
           |  |  |   ||| | |  ::|:|   |   | :|  : | ||     :|  |   :  |
Db     751 T--EVPPAVVKKPEPVAKSRDPSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDA 805

Qy     443 PTTPEELAPTTPEEP-----TPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKE 496
           |  |  :    : |     ||  | :|:|     | |:|  :   |  : |:| | : |||
Db     806 PAEPVNVPEPEKKTPVLAKKTPVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKE 865

Qy     497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTP 555
           |:|   :  :|  |    || :| |     | |   :  ||   |:|:   :|  |
Db     866 PSPKKAEPNSPVVP----PTPVKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLA 918

Qy     556 KGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPT-------TPKKPAP-----KEL 603
           |   |  ||: ||    :|   :||   |  ||||         :|||  |     :|
Db     919 K-KEPEKPKD-APKVAAKPRDPSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEA 974

Qy     604 APTTTKGPTS------TTSDKP-----APTTPKET------APTTPKEPAPTTPKKPAPT 646
             ||  | |           |:|     ||   |:|      ||  |::|   :||| ||
Db     975 PPTPVKNPVKKWKPPWEDDDEPAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPV 1031

Qy     647 TPETPPPTTSEVSTPTTTKEPTTIHKSP----DESTPELSA-EP---TPKALENSPKEPG 698
             : | |   ||  ||  | |    | |    || :  :|| ||   ||   : :| :|
Db    1032 AAK-PDPKIPEV-PPTPVKNPVKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPA 1089

Qy     699 V------------PTTKTPAAT------KPEMTTTAKDKTTERDLRTTPETT--TAAPK 737
                        ||:| |  :      ||: ||  ||   :      ||     ||||
Db    1090 TKPDSEAAADPVSGPTSKDPKLSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPK 1149

Qy     738 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
             |          |:  |          |     |            ||    |
Db    1150 KWKPVWDDDPDEPEADFTVPAPSKKPDTEDPADPLG---------GPKTKDPK------- 1193

Qy     798 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK--------- 848
            :||     || ::       |||:: :| | |||   ||   |:  ||
Db    1194 -LNKKAPAEKPTEK--------PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPE 1243

Qy     849 ---TTPTPRKMTSTMP 861
              | | |:|  :  |
Db    1244 ADFTMPAPKKPDTEDP 1259

RESULT   5
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacc
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Muc1, a mucin-like protein that is regulated by Mss10, is critical for pseud
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
```

C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosidase
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>

Query Match 14.6%; Score 789.5; DB 1; Length 1367;
Best Local Similarity 28.4%; Pred. No. 3.8e-27;
Matches 301; Conservative 104; Mismatches 483; Indels 171; Gaps 47;

```
Qy      35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
Db     236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy      87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
Db     296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy     145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
Db     345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy     195 KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAP 254
Db     405 SSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTES 453

Qy     255 TTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEP 307
Db     454 SSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSS 509

Qy     308 APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP 363
Db     510 APVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTP 568

Qy     364 ----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETA 418
Db     569 VTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSA 624

Qy     419 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKA 472
Db     625 PVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 679

Qy     473 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAP 531
Db     680 SAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 731

Qy     532 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTL 587
Db     732 SSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 790

Qy     588 KEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKK 642
Db     791 SAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESS 850

Qy     643 PAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP 697
Db     851 SAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYP 906

Qy     698 GVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT 757
Db     907 GSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT 966

Qy     758 --TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDR 811
Db     967 VPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA----- 1021

Qy     812 ATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPR 854
Db    1022 ---SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP- 1077

Qy     855 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHV 914
Db    1078 -SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGCSPKTVTTTVPCST 1122

Qy     915 FMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWML 974
Db    1123 GTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTL 1175

Qy     975 SPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1009
Db    1176 APSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207
```

RESULT 6
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

Query Match 14.5%; Score 784; DB 2; Length 1188;
Best Local Similarity 29.9%; Pred. No. 5.7e-27;
Matches 230; Conservative 59; Mismatches 359; Indels 122; Gaps 33;

```
Qy      13 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 68
Db     460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy      69 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 128
Db     511 APVG------------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy     129 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 183
Db     542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy     184 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 240
Db     600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy     241 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
Db     660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy     292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
Db     719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy     352 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 411
Db     769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy     412 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 458
Db     828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy     459 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 517
Db     886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy     518 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 563
Db     944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000

Qy     564 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 622
```

```
Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    623 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 681
Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 731
Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158


RESULT   7
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a muscle
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor

  Query Match             14.3%;  Score 776;  DB 2;  Length 2187;
  Best Local Similarity   27.7%;  Pred. No. 2.2e-26;
  Matches  288;  Conservative  122;  Mismatches  389;  Indels  240;  Gaps   49;

Qy      34 VTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKE---------TSLTVNKETT 84
Db     806 VQSPKVDPIMSDVTPTSPKKTSATAV-PK-------DTSATLSLKSVPAVTSLSPPKAPV 857

Qy      85 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPA 144
Db     858 APSNEATIVPTEIPTSLKNALAAATPKETLATSIPKVTSPS----PQKTPKSVSLKGAPA 913

Qy     145 LTT-----------------PKE-------PTPTTPKEPASTT---------PKEPTPTT 171
Db     914 MTSKKATEIAASKDVSPSQFPKEVPLLQHVPPTSPPKSPVSDTLSGALTSPPPKGP-PAT 972

Qy     172 IKSAPTTPKE-PAPTTTKSAPTTPK---------EPAPTTTKEPAPTTPKEPAPTTTKEP 221
Db     973 LAETPTYPKKSPKPAASKKTPATPSPEGVTAVPLEIPPCSKKAPKTAAPKESSATSSSKR 1032

Qy     222 APTT--TKSAPT--------------------TPKEPAPTTPKK-PAPTTPKEPAPTTP 257
Db    1033 APKTAVSKEIPSKGVTAVPLEISLPLKETSKSATPGEKSASSPKRSPKTAGPKE---TPP 1089

Qy     258 KEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTT--PKEPAP----TTPKEPAPTT 311
Db    1090 GGVTAVPPEISLPPKETPQNATPNESLAASSQKRSPKTSVPKETPPGGVTAMPLEIPSAP 1149

Qy     312 TKEPSPTTPK-----EPAPTTTKSAPTTTKEPAPTTT-KSAPTTPK--------EPSPTT 357
Db    1150 QKAPKTAVPKQIPTPEDAVTILAGSPLSPKKASKTAAPKEAPATPSVGVIAVSGEISPSP 1209

Qy     358 TKEPAPTTPKEPAPTTPKKPAPTT--PKE------------PAPTTPKEPAP-------T 396
Db    1210 KKTSKTAAPKENSATLPPKRSPKTAAPKETPATSSEGVTAVPSEISPSPPTPASKGVPVT 1269

Qy     397 TTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK-PAPTTPEELAPTTPE 455
Db    1270 LTPKGAPNALAE-SPASPKKVPKTAAPEETSTTP------SPQKIPKVAGPKEASATPPS 1322

Qy     456 EPTPTT--------PEEPAPTTPKAAAPNTPKEPAPTTPKE-PAPTTPKEPAPTTPKETA 506
Db    1323 KKTPKTAVPKETSAPSEGVTAVPLEIPPSPRKAPKTAAPKETPAPS--PEGATTAPVQIP 1380

Qy     507 PTTPKGTAPTTLKEPAPTTP-------------------KKPAPKELAPTTTKEPTS-- 544
Db    1381 PSPRKGSKKAGSKE-TPTTPSPEGVTAAPLEIPISSKKTSKMASPKETLVTPSSKKLSQT 1439

Qy     545 ----TTSDKPAPTTPKGTAPTTPKEPAPTTPKE-PAPTTPKGTAPTTLKEPAPTTPKKPA 599
Db    1440 VGPKETSLEGATAVPLEIPPSHKKAPKTVDPKQVPLTPSPK-DAPTTLAE-SPSSPKK-A 1496

Qy     600 PKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETP-----PPT 654
Db    1497 PKTAAPPSER-VTTVPPEKPA-TPQKASGTTASKVPVPAETQEVAVSSRETPVTPAVPPV 1554

Qy     655 TSEVSTPTTTK------EPTTIHKSPDESTPELSAEPTPKALENSPKE-PGVPTTKTPAA 707
Db    1555 KNPSSHKKTSKTIELKEAPATLPPSPTKSPKIPSSKKAPRT--SAPKEFPASPSIK-PVT 1611

Qy     708 TKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQ 767
Db    1612 T--SLAQTAPPSLQKAPSTTIPKENLAAPAV----LPVSSKSPAAPARASASLSPATAAP 1665

Qy     768 DTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTK 827
Db    1666 QTAPKEATTIPSCKKAAATETPIETSTAPSLEGAPKETSE----TSVSKVLMSSP----- 1716

Qy     828 APKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSK 887
Db    1717 -PKKASSSKRASTLP-------------ATTLPSLKEASVLS---------PTATSSGK 1752

Qy     888 LVEVNPKSEDAGGAEGETP 906
Db    1753 DSHISPVS-DACSTGTTTP 1770


RESULT   8
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episia
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precurso
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218;
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129;
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burc
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epith
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
```

A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated seque
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky, J.
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may genera
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <WR2
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associated
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu, H
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut agglu
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokine r
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region ar
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 125 c
partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively gly
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48146
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice form
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predicted
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

Query Match 12.4%; Score 672; DB 1; Length 1344;

Best Local Similarity 27.8%; Pred. No. 4.1e-22;
Matches 310; Conservative 78; Mismatches 472; Indels 256; Gaps 56;

```
Qy      77 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           |||   ||      |  |      :|| | || ||| :  |:  :: |: :: || ||:  :
Db      15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 166
           |    :||:|    |    ||                | |     ||||          :
Db      75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy     167 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 206
           | | :       : ||| |   ||| :|        |#| |   ||| :| |       |
Db     135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 252
           | |   ||| :|  ||   | ||| |   |  | |           :||| :   |
Db     191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy     253 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 289
             || |   | | |           ||| :  |       || |   |  |||
Db     248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy     290 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 342
             || |   | | |       || |   | | |      |:   |: || :|  ||
Db     308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy     343 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 391
           | ||| |   |:| :|  |       || |   | | |        || |   |  | | |
Db     368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy     392 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 447
               || |   | |||      |:  :| | |        |      || |   |  | | |
Db     425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy     448 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 487
               || |   | | |             |||  | | |    :||:|   |  | | |
Db     485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy     488 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 534
               || |   | | |     |     :|| |   | | |            :|||
Db     545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy     535 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 588
           ||       ||    :|||  | |      |:    || |   |  | |       :|| |
Db     601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy     589 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 632
             | | |           :|||  ||   | ||    :|||  | |       :|| |
Db     656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy     633 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 675
             | | |           :||| :  | ||     | | |   | :       :  :||
Db     715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy     676 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 724
               | |    : ::|     | :  | |         |:|   :||       |:  |
Db     773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy     725 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 775
           |  | :|        |:||    :|       |       |:  |       |  |    :|
Db     833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy     776 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 833
           :  |  ||  |      |:  :     | || |    |::  | | |   : ||
Db     889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy     834 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 885
             | | | |    | :|  |    :: |:  |    |:  |  :     |||   :  |
Db     947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy     886 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 945
           :  |   | :  | |:      |           ||    |  | : :      | : : |:
```

```
Db   1004 AHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy    946 ICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 981
                  | : |||             :||| ||
Db   1050 ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 1085


RESULT   9
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of chick
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1


  Query Match            12.4%;  Score 671;  DB 2;  Length 1151;
  Best Local Similarity  25.9%;  Pred. No. 3.9e-22;
  Matches  284;  Conservative  100;  Mismatches  419;  Indels  292;  Gaps   55;

Qy      14 TPKP-------PVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPS 64
           || |       |   :  :|   ||      | :    | :     || |||     ||
Db     113 TPPPSQGPAGTPPPSQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP- 170

Qy      65 LPPNSDTSKETSLTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDL 122
                :  | || | |   |  |  :    : :||    | :   :||  | :  ::
Db     171 -------ASAASPTVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EV 219

Qy     123 APTSKVLAKPTPKAETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPT 177
            | :  :||           |  |   :| | |:|  | |:   :|    :||  |:  |
Db     220 KPAAAAVAKEAKAVTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--T 277

Qy     178 TPKEPAPTT-----------TKSAPTTPKEPAPTTTK---------------------- 203
              :| | |            | ||:     | | ||
Db     278 AEAKPVPATASLMATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEA 337

Qy     204 EPAPTTPKEPAPTTTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPK 250
           :|||  | :  |      ||| |:  | |      |:  :       |  |:  | | |
Db     338 KPAPPPPPQQLPKAAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP- 396

Qy     251 EPAPTTPKEPTPTT----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKE 306
             |   ||  | ||    |::| |   : || ||  |    |:          || ||::
Db     397 -TAGAVPKASTGTTPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQ 448

Qy     307 P---APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKE 360
           |   | |||    |  |   | ||| : | | ::| |   |      |  || ||
Db     449 PVTKAATTTNATPPPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAAT 508

Qy     361 P----APTTPKEPAPTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAP 406
           |    | | |::| ||     |:  |  :| |          ||  |:|:   | | || |
Db     509 PGVPNAATDPQKPPPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIP 568

Qy     407 K-------EPAPTTPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA------ 450
           |       | | | :    | | |  : |     | |  ||  | |  |  ||
Db     569 KPVPLMALTPQPVTAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPPGLPRALAAAKLLG 627

Qy     451 -PTTP----------EEPTPTTP----EEPAPTTP---------KAAAPNTPKE------ 480
            |::|           | | :|      ||   |          ||:|    |
Db     628 LPSSPVASAMHAKVTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNG 687

Qy     481 ----PAPTTPKEPAPTTPKEPAPTTPKETAPT------------TPKGTAPT------- 516
               |      : ||      | | |   | |             :| | | |
Db     688 TLMAPMGAANTQMAPIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGA 747

Qy     517 ------------TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTP 555
                                                  |   |  |  : :|    | ||    |         |:|::      :|
Db     748 GGATQMSPMGAANTQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSP 807

Qy     556 KG--TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP 611
           |  | | ||    | ||| :|:|       | |||  |  | ||| :  |  :   ||: |
Db     808 MGAATPPQTPSVGAATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPP 859

Qy     612 ----TSTTSDKP-APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP------- 652
               |:||   | | :||: |   | || : |   | ||   |   :||: ||
Db     860 PMAATNTTQPPPMAASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPT 919

Qy     653 PTTSEVSTPTTTKEPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKP 710
           |      :||    |  : |||    |:   :|  |  |   :   ||| |  :||   |
Db     920 PPPPIPPSPTAQTSPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---P 975

Qy     711 EMTTTAKDKTTERDLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDT 769
             | |         |   |   |  || :   ||  ||  |:  :||  | | :    :
Db     976 IGVTEASPSADGARLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EA 1029

Qy     770 TPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP 829
           |          | | : :::  : :::   :    ::   |  ::| ::   |  |
Db    1030 AP---------TKAKRSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA--- 1077

Qy     830 KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLV 889
                  | |   :   ||    |  |        : || :|
Db    1078 ------------PAVGDGQQQMTPGAAQSVPP-------VTEAAVQ-------------- 1104

Qy     890 EVNPKSEDAGGAEGE 904
           |    :  | ||| |
Db    1105 EAAAAAAAAAGAERE 1119


RESULT  10
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3


  Query Match            12.2%;  Score 659.5;  DB 2;  Length 1229;
  Best Local Similarity  29.0%;  Pred. No. 1.3e-21;
  Matches  307;  Conservative   78;  Mismatches  339;  Indels  333;  Gaps   52;

Qy      33 KVTT-PDTSTTQ---HNKVSTSPKITTA-KPINPRPSLPPNSDT-------SKETSLTVN 80
           ||||  | |||      | ||:|:| |    || :     :  |       | | |  :
Db     119 KVTTSTDASTTNAPTTGKDSTTPEIITGIVVINSKSESVTDMSTTRFSTTLSPTTELLTS 178

Qy      81 KETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTT 140
             || | |  :|:| :||| |    :  |  || :  :  :  |:   ||    : |||
Db     179 PETLVSTDSSTST-EQTSPDNTTEIASPMETNTTTEATTTSVEPSVSTLAS---EDETTV 234

Qy     141 KGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPT 200
             | :|      ||      ||| :||| |             | :||| : |     ||
Db     235 TAIAEST------TTVIAEVSTTTEEPTTT------------AESTTKKSTT----KAPA 272

Qy     201 TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----KKPAPTTPKEPAPTT 256
           ||:|| ||| :|   |||:    ||| |  :| |   ||||    |  | | |   | |
Db     273 TTEEPTPTTTEE--VTTTEAETSTTTSSETSTEK---PTTPLIDNKIAGPATGK---PET 324
```

```
Qy    257 PKEP-TPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKE- 314
Db    325 THFPVTGTTPN--FDTATETPFVAKSEDKMTLSKTAATETTQQTTEVT-DGPEKETTKNV 381

Qy    315 ----PSPTTPKEPAPTTTKSAP-------TTTKEPAPT----TTKSAPTTPKEPSPTTTK 359
Db    382 SIEIPITTVPLVETTSTTSTASKESDGFHTTLKLKVTTADSDSTESATTVKPFNEETTTK 441

Qy    360 E---PAPT-------TPK------EPAPTTPKKPAP--------------------TTPK 383
Db    442 SHVVPKPTKKGTVKVTPKLELSFDEPTEIT-KAPHPGKLLEKKTYHFVLSDNFARYSEAK 500

Qy    384 E---------------PAPTTPKEPAP----TTTKKPAPTA--PKEPAPTT---PKETAP 419
Db    501 ENDDYNHLDYNHYREAKEPTTTEESSTTEEVTTTEEPANTGNPPTTENPTTTEQPTSTAE 560

Qy    420 TTPKKLTPTTPEKLA---PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPN 476
Db    561 STTTALPFTTEQTVTTEEPTTAEKSTATQ----KPTTTQESVST---EKTSTTKKA---S 610

Qy    477 TPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAP 536
Db    611 TTEE--PTTTDEPTTTT---ESSTTGKATTPELSTTSEETTTTELKITTE--------GS 657

Qy    537 TTTKEPTSTTSDKPAP----TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAP 592
Db    658 TTTEEPTTTAIFAEASTGIITTDEETTSTTSTTPEITSTKE--IVTESAITQTSVSVVES 715

Qy    593 TTP----------------------KKPAPKELAPTTTKGP---------------TS 613
Db    716 STPRQLPERWKAIVNKFKHNLEVLKEKKRLLKEKESTSTTGSDSSETTTVVAENIDEVTT 775

Qy    614 TTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPE--TPPPTTSE-VSTPTTTKEPTTI 670
Db    776 TEKEKVVQTTPITTEKSTTQEETTTTTTTTEKTTSKTTTEKPTTSESATTETTTSEPST- 834

Qy    671 HKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPE 730
Db    835 ----TEST---------------------TVDTSSATTEESSTAAETTTTSAE---TSE 865

Qy    731 TTTA-----------------------------------------APKMTKETATTTEKTTE 751
Db    866 TTTSESAAFITGESPENTALQSSSQKSEENESSAEKPGARRDFVPKKHKTTVKPAETTSA 925

Qy    752 SKITATTTQ-VTST---TTQDTTPFKITTLKTTTLAPKVT---TTKKTITTTEIMNKPEE 804
Db    926 VAASTTTTEPITTTEKSTTLETTPIEATTLNEVTGPAFVTGAPVDETTINTLELLSK--- 982

Qy    805 TAKPKDRATNSKATTPKPQ----------------KPTKAPKKPTSTKKPKTMPRVRKP 847
Db    983 -------INNTQISQPKPTDISKTDALSSLISGLIGSFTKAPMAPTI------------ 1022

Qy    848 KTTPTPRKMTSTMPELN-----------PTSRIAEA 872
Db   1023 HTTTDAAFVTATEASLNDGSDKKIIDEAQPTDEIRRA 1059
```

RESULT 11
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume i
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>

Query Match 11.7%; Score 633; DB 2; Length 489;
Best Local Similarity 31.2%; Pred. No. 7.5e-21;
Matches 149; Conservative 33; Mismatches 252; Indels 44; Gaps 7;

```
Qy    179 PKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAP 238
Db     39 PKQTPPYYYNAPPYYYKSPPP-----PSPSPP----------PPPYVHKYPPYYYKSPPP 83

Qy    239 TTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTP---KEPAPTAPKKPAPTT 295
Db     84 PSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPS--PSPPPPYYYKSPPPPSPSPPPPYY 141

Qy    296 PKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSP 355
Db    142 YKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPP-------------PPSPSPPPP 187

Qy    356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPK 415
Db    188 YYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYYKSPPPPSPS 247

Qy    416 ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAP 475
Db    248 PPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPP 307

Qy    476 NTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 535
Db    308 PSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYVYKSPPPPSPSPPPPYYY- 366

Qy    536 PTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP 595
Db    367 ----KSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSP 422

Qy    596 KKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPP 653
Db    423 SPP-----PPYVYKSPPPPSPSPPPPYYYKSPPPPSPSPPPPYYYKSPPPPSPSPPP 475
```

RESULT 12
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840

```
A;Map position: 2q31-2q31

  Query Match           11.7%;  Score 632.5;  DB 2;  Length 7962;
  Best Local Similarity  22.9%;  Pred. No. 1.1e-19;
  Matches  295;  Conservative  105;  Mismatches  453;  Indels  437;  Gaps   60;

Qy      10 KKKPTP----KPPV---VDEAGSGLDNGDFKVT-----------TPDTSTTQHNKVSTS 50
           || | |    | ||   :::    ::    :||           |  :  |   :
Db    5837 KKAPPPPKVIKKPVIEKIEKTSRRMEEEKVQVTKVPEVSKKIVPQKPSRTPVQEEVIEVK 5896

Qy      51 -PKITTAKPINPRPSLPPNSDTSKETSLT---VNKETTVETKETTTTNKQTSTDGKEKTT 106
            | : | | :     :   | | :| |:|   | ||   | |     :|:
Db    5897 VPAVHTKKMVISEEKMFFASHTEEEVSVTVPEVQKEIVTEEKIHVAVSKRVE-------P 5949

Qy     107 SAKETQSIEKTSAKDLAPT--SKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPAS--- 161
             |  :  || : :::||    | :  | ||     | | :   |:| |   ||||:
Db    5950 PPKVPELPEKPAPEEVAPVPIPKKVEPPAPKVPEVPKKP-VPEEKKPVPVPKKEPAAPPK 6008

Qy     162 --TTPKEPTPTTIKSAPTTPKEPAP-----------TTTKSAPTTPKEPAPTTTKEPA- 206
             ||:| |      |   |: ||            |  |    | :| :|     ||
Db    6009 VPEVPKKPVPEEKIPVPVAKKKEAPPAKVPEVQKGVVTEEKITIVTQREESP----PPAV 6064

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP-KEPAPTTPKEPTPTT- 264
           |  ||:  |   | | |   :  |  || ||   |||| |      | |   | | |
Db    6065 PEIPKKKVPEERK-PVPRKEEEVPPPPKVPA--LPKKPVPEEKVAVPVPVAKKAPPPRAE 6121

Qy     265 ------------------------------------------------------------P 265

Db    6122 VSKKTVVEEKRFVAEEKLSFAVPQRVEVTRHEVSAEEEWSYSEEEEGVSISVYREEEREE 6181

Qy     266 KEPAPTTK-------------------------EPAPTT----------PKEPA------ 284
           :| |  |:                         |||  |          || ||
Db    6182 EEEAEVTEYEVMEEPEEYVVEEKLHIISKRVEAEPAEVTERQEKKIVLKPKIPAKIEEPP 6241

Qy     285 ----PTAPKK-------PAPTTPKE--PAPTTPKEP-APTTTKEPSPTTPKE----PAPT 326
               | ||||       |||   ||  | |  |:||  |   |:  |   |     ||
Db    6242 PAKVPEAPKKIVPEKKVPAPVPKKEKVPPPKVPEEPKKPVPEKKVPPKVIKMEEPLPAKV 6301

Qy     327 TTKSAPTTTKEP--APTTTKSAP---TTPKEPSPTTTKE-------------PAPTT--- 365
           | |    | :|      | | ||     |:||     :|              ||  |
Db    6302 TEKHMQITQEEKVLVAVTKKEAPPKARVPEEPKRAVPEEKVLKLKPKREEEPPAKVTEFR 6361

Qy     366 -------------PK-EPAP----------------------------------------- 371
                        || || |
Db    6362 KRVVKEEKVSIEAPKREPQPIKEVTIMEEKERAYTLEEEAVSVQREEEYEEYEEYDYKEF 6421

Qy     372 ------------------------------TTPKKPAPT--TPKEPAPTTPKEPAPTTTK 399
                                           | |:|| |    |:|| || || |     |
Db    6422 EEYEPTEEYDQYEEYEEREYERYEEHEEYITEPEKPIPVKPVPEEPVPTKPKAPPAKVLK 6481

Qy     400 KPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK-------LAPTTPEKPAPTTPEELAPT 452
           |  |   | | |  ||:  |  ||   |  |:|        :: |  ||   | |   |
Db    6482 KAVP-EEKVPVP-IPKKLKPPPPK--VPEEPKKVFEEKIHISITKREKEQVTEP---AAK 6534

Qy     453 TPEEPTPTTPEEPAPTTPKAAAP--NTPKEPAPTTPKEPA------------------- 490
            | :|     ||  |   |  ||    |: |    |:| |
Db    6535 VPMKPKRVVAEEKVPVPRKEVAPPVRVPEVPKELEPEEVAFEEEVVTHVEEYLVEEEEEY 6594

Qy     491 ------PTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTS 544
                 | :|  |  | :  |  |:   |   |:| |   || ||     |   |:|
Db    6595 IHEEEEFITEEEVVPVIPVK-VPEVPRKPVPEE-KKPVPVPKKKEAPPAKVPEVPKKPEE 6652

Qy     545 TT------SDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPK 596
                  :||    |    |  ||:| |   | | |   |  ||     ||  |  ||
Db    6653 KVPVLIPKKEKP----PPAKVPEVPKKPVP-EEKVPVPVPKKVEAP-----PAKVPEVPK 6702

Qy     597 KPAPKEL----APTTTKGPTSTTSDKPAPTTPKETAPT-TPKE---PAPTTPKKPAPTTP 648
           || |::     ||   : | :   : |    |:|  ||  ||:   | |  |||  |
Db    6703 KPVPEKKVPVPAPKKVEAPPAKVPEVPKKLIPEEKKPTPVPKKVEAPPPKVPKKREPVPV 6762

Qy     649 ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAAT 708
                 |  ||           :    :|  ||   |  |:  |  |:|  :|  :
```

```
Db    6763 PVALPQEEEVLFEEEIVPEEEVLPEEEEVLPE-EEEVLPEEEEVLPEEEEIPPEEEEVPP 6821

Qy     709 KPEMTTTAKDKTTERDL--RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
           : |      ::   | ::      |:    ||          :| || |:     :
Db    6822 EEEYVPEEEEFVPEEEVLPEVKPKVPVPAP-----VPEIKKKVTEKKVVIPKKEEAPPAK 6876

Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPT 826
             | |:   :   : ||    :: :  |:  :|||              :|
Db    6877 VPEVPKKVE--EKRIILPK----EEEVLPVEVTEEPEE---------------EPISEE 6914

Qy     827 KAPKKPTSTKK-----PKTMPRVRK---PKT-TPTPRKMTSTMPELNPTSRIAEAMLQTT 877
           : |::| | ::      |  :| | |   |:  || |:|: :  |    : :| | :
Db    6915 EIPEEPPSIEEVEEVAPPRVPEVIKKAVPEAPTPVPKKVEA--PPAKVSKKIPEEKVPVP 6972

Qy     878 TRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIIN 937
            :  : | :|: ||                          |:  |:   |  || :  :
Db    6973 VQKKEAPPAKVPEV-----------------------PKKVPEKKVL--VPKKEAV-- 7003

Qy     938 PMLSDETNICNGKPVDGLTTLRNGTLVAFR 967
                       |  | | |     ||||
Db    7004 -----------PPAKGRTVLEEKVSVAFR 7021


RESULT  13
C84672
hypothetical protein At2g27380 [imported] - Arabidopsis thaliana
C;Species: Arabidopsis thaliana (mouse-ear cress)
C;Date: 02-Feb-2001 #sequence_revision 02-Feb-2001 #text_change 02-Feb-2001
C;Accession: C84672
R;Lin, X.; Kaul, S.; Rounsley, S.D.; Shea, T.P.; Benito, M.I.; Town, C.D.; Fujii, C.Y
M.; Koo, H.; Moffat, K.S.; Cronin, L.A.; Shen, M.; VanAken, S.E.; Umayam, L.; Tallon,
euss, D.; Nierman, W.C.; White, O.; Eisen, J.A.; Salzberg, S.L.; Fraser, C.M.; Venter
Nature 402, 761-768, 1999
A;Title: Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana.
A;Reference number: A84420; MUID:20083487
A;Accession: C84672
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-761 <STO>
A;Cross-references: GB:AE002093; NID:g5306260; PIDN:AAD41992.1; GSPDB:GN00139
C;Genetics:
A;Gene: At2g27380
A;Map position: 2


  Query Match           11.7%;  Score 632;  DB 2;  Length 761;
  Best Local Similarity  31.1%;  Pred. No. 1.3e-20;
  Matches  217;  Conservative   44;  Mismatches  354;  Indels   82;  Gaps   33;

Qy     132 PTPKAETTTKGPALTTP---KEPTPT-TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTT 187
           | |  :  |  | :  |   | ||||  :|        | ||||        | :  || |
Db      69 PPPIQKPPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPT 128

Qy     188 KSAPTTPKEPAPTTTKEPAP--TTPKEPAPTTTKEPAPTTTKSAPTTP----KEPAPT-- 239
           | |  |    |   | | |  : | :| |    :  || | | |  |    | | ||
Db     129 YSPPIYP----PPIQKPPTPSYSPPVKPPPV---QMPPTPTYSPPIKPPPVHKPPTPTYS 181

Qy     240 TPKKP---APTTPKEPAPTTP----KEPTP--TTPKEPAPTTKEPAPT-TPKEPAPTAPK 289
           | ||    | ||     |  |    | |||  : | :| |  | | || :|    |   |
Db     182 PPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHK 241

Qy     290 KPAPT-TPKEPAPTTPKEPAP---TTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 345
           | |  :|    |   | | |      | |   ||  |  :     |   | | || :
Db     242 PPTPIYSPPIKPPPVHKPPTPIYSPPVKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPP 301

Qy     346 APTTPKEPSPTTTKEP---APTTPKEPAPT--TPKKPAPTTPKEPAPTTPKEPAPTTTKK 400
           : | :  || | |     |   | | ||  | || |  |  | :| | |   |
Db     302 VKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPVKP-PPVHKP 360

Qy     401 PAP--TAPKEPAPT----TPKETAPTTP----KKLTPTTPEKLAPTTPEKPAPTT---PE 447
           | |  : | :| |     ||  :  |  |    |  |||    :  |   :||   |  |
Db     361 PTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPI 420
```

```
Qy    448 ELAPTTPEEPTPTTPEEPAPT----TPKAAAP-NTPKEPAPTTPKEPAPTTPKEPAPTTP 502
Db    421 KLPPVKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTP 480

Qy    503 KETAPTTPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
Db    481 TYSPPVQP----PPVQKPPTPTYSPPVKPPPIQKPPT----PTYSPPIKPPPVKPPTPTY 532

Qy    561 TTPKEPAPTTPKEPAPT-TPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKGPTSTTSD 617
Db    533 SPPIKP-PPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPP 591

Qy    618 KPAPTTPKETAPTTPKEPAPTTPKKPAPT-TPETPPPTTSEVSTPTTTK--EPTTIHKSP 674
Db    592 VHKPPTPTYSPPIKP----PPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPP 647

Qy    675 DESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA 734
Db    648 ---TPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPP-TPTYSPPVKPPPVQVPPTPTYS 703

Qy    735 APKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP 771
Db    704 PPVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPP 740
```

RESULT 14
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/2; 3504/1

Query Match 11.6%; Score 629; DB 2; Length 3507;
Best Local Similarity 24.1%; Pred. No. 7.2e-20;
Matches 252; Conservative 158; Mismatches 416; Indels 218; Gaps 45;

```
Qy     10 KKKPTPKPPVVDEAGSGLDN-GDFKVTT----PDTSTTQHNKVSTSPKITTA-------K 57
Db   1942 KKQPNREKIEIDEENSSSSNSGQEKPTTKGIVSSTSATSSESTTAEPHVTTSISSTTSTK 2001

Qy     58 PINPRPSLPPN---SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQS 113
Db   2002 DMTSSKS-PENVTMSSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEV 2060

Qy    114 IEKTSAKDLAP----TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
Db   2061 ITESSVKSTTPKEESSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKS 2119

Qy    170 TTIKS-APTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKS 228
Db   2120 TVLSSEAPVTSTSPTEVHT-SSETKPSLSASSTTGDTNSTTPSTSSLASVKSTSAPEGTS 2178

Qy    229 APTTPKEPAPTTPKKPAPTTPK-EPAPTTPKEPTPTTPKEPAPTTKEP-APTTPKEPAPT 286
Db   2179 ASVAPVKLSSLSPDVSQPSTKTFDATESSTVQASETSSGTSVKSTSEPESHVTKLSITSS 2238

Qy    287 APKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP--KEPAPTTTKSAPTTTKEPAPTTTK 344
Db   2239 NPSSSVPVTSPKSTPTVPE-----STEQPTSTTPSGQSLTPMNSNSEVLTTSEPHVLSSS 2293

Qy    345 SAP------TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT- 397
Db   2294 LSPDVSQSSTTPNNLSESSTVE----TPKTSSEVSLNSEEPSTTEAPTTLSPDILSTTTN 2349

Qy    398 --------------------TKKPAPTAPK----------EPAPTTPKETAPTTPKKLTP 427
Db   2350 NLSQSSTVSTEDRSEISSENSEKPT-SAPELVTSSVTHVASSSPDVPTES--SEPDDLTG 2406

Qy    428 TTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP-----------APTTPKAAAPN 476
Db   2407 SSTENIPEASSKQTISSTPTPDTTTASEEPTKSTSMSPDLSTTSNVLSESSTTPESSS-- 2464

Qy    477 TPKEPAPTT-----------PKEPAPT---------TPKEPAP-----TTPKETAPTTP 510
Db   2465 --KSPVSSSTEGISVVTSTEFSKVPESTISSVLEEDLTKTTPSPILEETTTASETSEPLT 2522

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPA---------PT-------- 553
Db   2523 EDSLTVSVRIHELTTSSENVPKESESTTTSSESSKPSQEPAGILTSTVVVPTSSVSLITA 2582

Qy    554 ---------TP--KGTAP--TTPK-------EPAPTTPKEPAPTTPKGTAPTTLKEPAPT 593
Db   2583 SEIEAITSNTPFKQGRTPITTSPKSLVKSTTSPSTVTSSEPSESTKRTTVSTTVSTTTPT 2642

Qy    594 TPKKPAPKELAPTTTKGPT-STTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPP 652
Db   2643 EETTTSESLILTAAPSKPTESTTESSEAPTTPAKTSETKPSNVSSTSRKSTENVETSTSQ 2702

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTP--ELSAEPTPKALE--NSPKEPGVPTTKTPAAT 708
Db   2703 SGSLESSTMSSTS------SEPETNAPAVTVSSEASSTTLEENSSTSSP----TSSEASV 2752

Qy    709 K-----PEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTT---- 759
Db   2753 KLSSLFPESITSEAVTVSSR----APAEITMSSESHREISTVSSEPSEPEIPLSTTVSPN 2808

Qy    760 -----------QVTSTTTQDTTP-FKITT-----LKTTTLAPKVTTTKKTITTTEIMNKP 802
Db   2809 VVTASSIPSEEPILSSVTSSSTPRVRLITGTPDDLIVSVTVPSHGNRRQNITASSV---P 2865

Qy    803 EETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK---TTPTPRKMTST 859
Db   2866 SNSTSPIILPSES-LTTPQPPPTTTTTAKPATTSGKRGPPSIQPPAEMFTTPAP------ 2918

Qy    860 MPELN-----PTSRIAEAMLQTTT 878
Db   2919 PPPSNGGYGEETNQEEEQVTSTTT 2942
```

RESULT 15
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp
A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNoppl80
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein

Query Match 11.6%; Score 626.5; DB 2; Length 990;
Best Local Similarity 26.2%; Pred. No. 2.8e-20;
Matches 253; Conservative 132; Mismatches 393; Indels 187; Gaps 45;

```
Qy       3 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 62
Db      59 DAKKRKPPANGLPKKKSAKESSSSEDSSSEEDEPPAKKRAQ----PAGGKKPVVKAVQPK 114

Qy      63 PSLPPNSDTSKETSLTVNKETTVE--TKETTTTNK----QTSTDGKEKTTSAKETQSIEK 116
Db     115 KAKSSSEDSSDESD---SEEETKKPPAKRPAQTPKVAAVKTPTQKKAKSSSS-ESSSSED 170

Qy     117 TSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKEPTPTTI 172
Db     171 EASKKKQPVIKV----PPKQAVVKAGLASNNGKTADSSSSEDSDSPPAKKTAATKTPPT- 225

Qy     173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTT-TKEPAPTT-----T 226
Db     226 --KPATAAKPQAKKTAGKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVPPPTSVSKKKT 283

Qy     227 KSAPTTPKEPAPT----------TPKKPAPTTPKEPAPTTP----KEPTPTTPKEPAPTT 272
Db     284 LSQPGTKAKPESSDSSDSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSESDSSS 343

Qy     273 KEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTP----KEPAP----------------TT 311
Db     344 EDEKKSSVKLGVKAAPKK-APAAP--DAKSTPVAAAKKSAPAKKASSSSDSDSSSNEETT 400

Qy     312 TKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA- 370
Db     401 TKPAAKTTPAKSAATPTSKTPTNGK---ATPTSKTPAKPGTPKTSTAKKDSSSSDSSDSS 457

Qy     371 ---PTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPA-PTTPKETA------- 418
Db     458 SDEETTTKPAAKTTPAKSAATPTSKTP---TNSKATPTSKKTPAKPGTPKTSAAKKDSSS 514

Qy     419 -----PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP-KA 472
Db     515 SDSSDSSSDEKKTPA--KRAAKTTPAKPA-------AKTTPAKPA-------AKTTPAKP 558

Qy     473 AAPNTPKEPAPT----TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKEPAPTTPK 527
Db     559 AAKSTPGKQVPTKKESSSSDSSDSSSEDEKKSSAKPAVKTTPGKATS-------KPVVAS 611

Qy     528 KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT--PKEPAPTTPKGTAPT 585
Db     612 KPVPAK------KASSSSDSDSSEEETTKTTKPLTKLSPAVKTLPPKKAESSSDSSSDSD 665

Qy     586 TLKEPAPTTPKKPAPKELAPTTTKGP--------TSTTSDKPAPTTPKETAPT--TPKEP 635
Db     666 SEKK---TKPAKPPAKSATPVNTKAPAQNKASKASCSDSDSSSEEEGKSKQPTGKSPAAK 722

Qy     636 APTTPKK-PAPTTPETPPPTTSEVSTPTTTKE-PTTIHKSPDESTPELSAEPTPKALENS 693
Db     723 ATAPPKKNPVAVNKDKPSSSSSSDSSGDDEKQKPKQAAAAKDVKQGAKAAKPTPKKAASS 782

Qy     694 PKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESK 753
Db     783 SSE---DSSSDEDVSKAKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGG 835

Qy     754 ITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKD 810
Db     836 KNTSTTKIANST-----------------PKAAAAECSEESSSSEDEGKANGTSGKRKR 877

Qy     811 RATNS---KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPT 866
Db     878 ESTGNAECEAVTPE--------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP- 928

Qy     867 SRIAE 871
Db     929 -RMAD 932
```

Search completed: April 26, 2002, 16:19:40
Job time: 370 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:19:50 ; Search time 47.4 Seconds
(without alignments)
783.575 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1212
Perfect score: 5416
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....SPIDTVFTRCNCEGKTFFFK 1013

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1168	21.6	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	942	17.4	1664	1	SLP1_CLOTM	Q06852 clostridium
3	789.5	14.6	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	12.0	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	556.5	10.3	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	10.3	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	10.2	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	9.8	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	9.8	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	9.8	865	1	CPN_DROME	Q02910 drosophila
11	518.5	9.6	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	9.4	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	9.3	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	9.2	1970	1	RPB1_MOUSE	P08775 mus musculu
15	497	9.2	831	1	NFH_RAT	P16884 rattus norv
16	493.5	9.1	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	9.0	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488	9.0	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	8.8	267	1	EXTN_MAIZE	P14918 zea mays (m
20	471.5	8.7	5376	1	ZAN_MOUSE	O88799 mus musculu
21	468.5	8.7	634	1	HWP1_CANAL	P46593 candida alb
22	467	8.6	817	1	VRP1_YEAST	P37370 saccharomyc
23	463	8.5	2142	1	BAT2_HUMAN	P48634 homo sapien
24	454	8.4	797	1	VGLX_HSVEB	P28968 equine herp
25	454	8.4	1161	1	YJ9P_YEAST	P47179 saccharomyc
26	448.5	8.3	670	1	VG50_HSVI1	Q00130 ictalurid h
27	442.5	8.2	751	1	FP1_MYTGA	Q27409 mytilus gal
28	439.5	8.1	1083	1	T2D3_HUMAN	O00268 homo sapien
29	432.5	8.0	439	1	XP2_XENLA	P17437 xenopus lae
30	432	8.0	3164	1	TEGU_HSV11	P10220 herpes simp
31	426	7.9	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
32	424.5	7.8	1125	1	MAP4_MOUSE	P27546 mus musculu
33	424	7.8	307	1	SGS3_DROME	P02840 drosophila
34	424	7.8	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
35	422.5	7.8	1794	1	YAV1_SCHPO	Q10172 schizosacch
36	422.5	7.8	3421	1	TEGU_HSVEB	P28955 equine herp
37	419.5	7.7	2476	1	ZAN_PIG	Q28983 sus scrofa
38	419.5	7.7	2774	1	MAPA_RAT	P34926 rattus norv
39	419	7.7	1185	1	DRPL_HUMAN	P54259 homo sapien
40	415	7.7	1229	1	N121_HUMAN	Q9y2n3 homo sapien
41	411.5	7.6	907	1	VGP3_EBV	P03200 epstein-bar
42	410	7.6	1183	1	DRPL_RAT	P54258 rattus norv
43	407.5	7.5	3256	1	KI67_HUMAN	P46013 homo sapien
44	405.5	7.5	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
45	405	7.5	1411	1	TCOF_HUMAN	Q13428 homo sapien

ALIGNMENTS

```
RESULT   1
MUC2_HUMAN
ID   MUC2_HUMAN     STANDARD;      PRT;  5179 AA.
AC   Q02817; Q14878;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN   MUC2 OR SMUC.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Intestine;
RX   MEDLINE=94132002; PubMed=8300571;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT   "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT   Identification of the amino terminus and overall sequence similarity
RT   to prepro-von Willebrand factor.";
RL   J. Biol. Chem. 269:2440-2446(1994).
RN   [2]
RP   SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC   TISSUE=Colon;
RX   MEDLINE=93016075; PubMed=1400449;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA   Kim Y.S.;
RT   "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT   both upstream and downstream of its central repetitive region.";
RL   J. Biol. Chem. 267:21375-21383(1992).
RN   [3]
RP   SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX   MEDLINE=91358717; PubMed=1885763;
RA   Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA   Petersen G.M., Kim Y.S.;
RT   "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT   and polymorphism.";
RL   J. Clin. Invest. 88:1005-1013(1991).
CC   -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC       OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC       PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC       AGENTS AT MUCOSAL SURFACES.
CC   -!- SUBUNIT: MULTIMERIC.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC       BRONCHUS, CERVIX AND GALL BLADDER.
CC   -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC       INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC       VARIES AMONG DIFFERENT ALLELES.
CC   -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC       OF SILKWORM HEMOCYTIN.
CC   -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC   ---------------------------------------------------------------------
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL       1     20       POTENTIAL.
FT   CHAIN       21   5179       MUCIN 2.
FT   DOMAIN    1401   1747       APPROXIMATE REPEATS.
FT   REPEAT    1401   1416       1.
FT   REPEAT    1417   1432       2.
FT   REPEAT    1433   1448       3.
FT   REPEAT    1449   1464       4.
FT   REPEAT    1465   1471       5.
FT   REPEAT    1472   1478       6.
FT   REPEAT    1479   1494       7A.
FT   REPEAT    1495   1517       7B.
FT   REPEAT    1518   1533       8A.
FT   REPEAT    1534   1556       8B.
FT   REPEAT    1557   1572       9A.
FT   REPEAT    1573   1596       9B.
FT   REPEAT    1597   1612       10A.
FT   REPEAT    1613   1635       10B.
FT   REPEAT    1636   1651       11A.
FT   REPEAT    1652   1675       11B.
FT   REPEAT    1676   1683       12.
FT   REPEAT    1684   1699       13.
FT   REPEAT    1700   1715       14.
FT   REPEAT    1716   1731       15.
FT   REPEAT    1732   1747       16.
FT   DOMAIN    4815   4886       VWFC 1.
FT   DOMAIN    4924   4991       VWFC 2.
FT   DOMAIN    5075   5160       CTCK.
FT   DISULFID  5075   5122       BY SIMILARITY.
FT   DISULFID  5089   5136       BY SIMILARITY.
FT   DISULFID  5098   5152       BY SIMILARITY.
FT   DISULFID  5102   5154       BY SIMILARITY.
FT   DISULFID     ?   5159       BY SIMILARITY.
FT   CARBOHYD   163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD  5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT  1351   1351       H -> L (IN REF. 3).
FT   CONFLICT  1412   1412       T -> S (IN REF. 3).
FT   CONFLICT  1449   1449       L -> P (IN REF. 3).
FT   CONFLICT  1504   1504       M -> T (IN REF. 3).
FT   CONFLICT  4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE  5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;
```

```
  Query Match             21.6%;  Score 1168;  DB 1;  Length 5179;
  Best Local Similarity   32.5%;  Pred. No. 1.4e-39;
  Matches  363;  Conservative   53;  Mismatches  448;  Indels  254;  Gaps   40;

Qy      21 DEAGSGLDNGDFK----------------VTTPDTSTTQH-NKVSTSPK----------- 52
           |   || |:|| :                |  |  |  ||  ||
Db    1312 DHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQF 1371

Qy      53 ------------------------ITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
                                    |||  |    || || : |:   : | :  ||
Db    1372 GNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTT---- 1427

Qy      89 ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTP 148
            ||||   |:|      ||:                ||:      |:|   ||| |  |||
Db    1428 -TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPISTTTPPPTTTP 1470

Qy     149 KEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------APTTTKSAP-TTP 194
             ||      |||   | :||   | |||  | | |||  |           ||||  | | ||
Db    1471 SPPTTTPSPPTTTPSPPTTTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTT 1530

Qy     195 KEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPKEPAPTTPKKPA 245
             | ||||  |  |||  |         ||||  | ||||  : |  |||  |  |||  |
Db    1531 TTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPT 1590

Qy     246 PTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAPKKPAPTTPKE 298
             ||   |  |||  || ||   |  ||  |  |||  |         ||    | |||
Db    1591 ITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTT 1650

Qy     299 PAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP---TTPKEPSP 355
             | |||   |  |||  |  ||   | |||| |:| ||    |||| : |   |||  |
Db    1651 PPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTPSPPTTTMTTPSPTTTPSSPI- 1709

Qy     356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTP- 414
            |||  |: ||   | |||   |:|||   |  ||     ||||  |  | |  |: | |
Db    1710 TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPT 1769

Qy     415 ----KETAPTTP-------------------------------------------KKLTP 427
                | ||||                                             : |
Db    1770 FSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYP 1829

Qy     428 ------------------------------------------------------TTPEKLAPTT 437
                                                                 | |  :  ||
Db    1830 DVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTMTTTT 1889

Qy     438 PEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKEP-----AP 491
            | | | |   :  ||   |||| |  :   |||          | ||  : |
Db    1890 TENPTPPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT 1949
```

```
Qy     492 TTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKP 550
           ||   | || |  :|  |||  |  |    | ||  : |    :  |||  || | :
Db    1950 TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQ 2009

Qy     551 AP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 609
           | |||  |  |    | ||  : |  |||  |  |    | ||  : |    :  |||
Db    2010 TPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV 2068

Qy     610 GPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT 668
           || | :    | |||  |  |    | ||  : |  ||| |   | :   ||| |: ||
Db    2069 TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPT 2127

Qy     669 TIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPAATKPEMTTTAK 717
           |   :|  :|  ::  |||   :     |            || |:||  |    |||
Db    2128 T---TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2184

Qy     718 DKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTTQVTSTTTQDT 769
           |    :   |||  |||  |  |    :| |||  || : :| | |  | | |  |
Db    2185 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP-TGTQTPTT 2243

Qy     770 TPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKA 828
           ||  |||  |||: |  | |  :| ||| |     |  |    | :  |||     |
Db    2244 TP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2299

Qy     829 P-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAMLQTTTR-PNQT 883
           |   || |: | | |       ||||    :  |   |   |: :      | |: |  |
Db    2300 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT 2359

Qy     884 PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 921
           | :    | |     |    :||     |      |||
Db    2360 PITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392
```

```
RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL        1     28       POTENTIAL.
FT   CHAIN        29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN       36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT       36    191       1.
FT   REPEAT      207    363       2.
FT   REPEAT      409    565       3.
FT   REPEAT      607    763       4.
FT   DOMAIN      771   1377       APPROXIMATE TANDEM REPEATS OF
FT                                T-P-S-D-E-P.
FT   DOMAIN     1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN     1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN     1495   1565       SLH 2.
FT   DOMAIN     1566   1625       SLH 3.
FT   DOMAIN     1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;
```

```
  Query Match             17.4%;  Score 942;  DB 1;  Length 1664;
  Best Local Similarity   33.0%;  Pred. No. 5e-31;
  Matches  276;  Conservative   83;  Mismatches  312;  Indels  166;  Gaps   43;

Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |   ||||:          ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :   ||:|| || |    ||   ||     || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     302 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPTTTKSAPTTPKEPSPTT 357
           :    |:  || : ||:|: ||:|| || | |  || : || |:             ||
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS----------DEPTP 968

Qy     358 TKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPAPTTPKE 416
           : || |:   || |:      ||   || |: ||:|| || |    ||   || |:
Db     969 SDEPTPS--DEPTPS----DEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSD--- 1019

Qy     417 TAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAP 475
             ||  ::  ||  ::| |:      ||  :|||| |||| ||   :  |
Db    1020 ---------EPTPSDE--PTPSDEPTPSD----EPTPSDEPTPSETPEEPIPTDTPSDEP 1064

Qy     476 NTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPK 532
           || |:    |:  ||   || |: ||:|  ||      ||   || |:   :| |
Db    1065 TPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS 1122

Qy     533 ELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA 591
           :  || : ||| | | ::| ||      ||   || |:   || |:      ||   ||
Db    1123 D-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS----DEPTPSDEPT 1175

Qy     592 PT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKEPAPT-TPKKPAPT 646
           |: ||::|      || |     | ||:| |:   || :  ||   || |: ||::| ||
Db    1176 PSETPEEP-----IPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETPEEPIPT 1229

Qy     647 TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPA 706
             : || |:   || : |||      || ||  | ||||     :|:|| :||
Db    1230 DTPSDEPTPSD--EPTPSDEPT----PSDEPTP--SDEPTP---SETPEEP-IPTDTPSD 1277

Qy     707 ATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
           |    | |: |  |  |  : | : : |      |  :| |  |   |   | :
Db    1278 EPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPS 1337

Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ-KP 825
           : ||   |                    | :  | |  | |   ||  :  |
Db    1338 DEPTPSDEPT----------------------PSDEPTPSDEPTPSDEPTPSDEPTP 1372

Qy     826 TKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTS--------------TMPELNPTSRIA 870
           :: |::|| |  |        | ||||     |               | |   |||
Db    1373 SETPEEPTPTTTP-----TPTPSTTPTSGSGGSGGSGGGGGGGGGGTVPTSPTPTPTS--- 1424
```

```
Qy     871 EAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLP 927
Db    1425 --------KPTSTPAPTEIE-EPTPSDVPGAIGGEHRAYLRGY-------PDGSFRP 1465

RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL        1     21       POTENTIAL.
FT   CHAIN        22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN      210   1367       SER/THR-RICH.
FT   CARBOHYD    817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;

  Query Match             14.6%;  Score 789.5;  DB 1;  Length 1367;
  Best Local Similarity   28.4%;  Pred. No. 4.6e-25;
  Matches  301;  Conservative  104;  Mismatches  483;  Indels  171;  Gaps   47;

Qy      35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
Db     236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy      87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
Db     296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy     145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
Db     345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy     195 KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAP 254
Db     405 SSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTES 453

Qy     255 TTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEP 307
Db     454 SSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSS 509

Qy     308 APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP 363
Db     510 APVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTP 568

Qy     364 ----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETA 418
Db     569 VTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSA 624

Qy     419 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKA 472
Db     625 PVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 679

Qy     473 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAP 531
Db     680 SAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 731

Qy     532 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTL 587
Db     732 SSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 790

Qy     588 KEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKK 642
Db     791 SAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESS 850

Qy     643 PAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP 697
Db     851 SAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESPSTGT-TVTPSSSKYP 906

Qy     698 GVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT 757
Db     907 GSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT 966

Qy     758 --TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDR 811
Db     967 VPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA----- 1021

Qy     812 ATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPR 854
Db    1022 ---SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP- 1077

Qy     855 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHV 914
Db    1078 -SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGCSPKTVTTTVPCST 1122

Qy     915 FMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWML 974
```

```
          | | :   |        ::    |  ||   ||   | ||    |       |   |
Db   1123 GTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTL 1175

Qy    975 SPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1009
          :| :| :||        :|: | |  | |    |  |:|
Db   1176 APSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207


RESULT  4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
```

```
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23       POTENTIAL.
FT   CHAIN        24   1255       MUCIN 1.
FT   DOMAIN       24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186       POTENTIAL.
FT   DOMAIN     1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151       SEA.
FT   CARBOHYD    957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                                ISOFORM).
FT   VARSPLIC   1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                ISOFORM).
FT   VARSPLIC   1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2       T -> A (IN REF. 11).
FT   CONFLICT    134    134       P -> Q (IN REF. 9).
FT   CONFLICT    154    154       P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021       S -> T (IN REF. 3).
FT   CONFLICT   1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;
```

Query Match 12.0%; Score 651; DB 1; Length 1255;
Best Local Similarity 27.8%; Pred. No. 1.3e-19;
Matches 295; Conservative 72; Mismatches 460; Indels 234; Gaps 51;

```
Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||          | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTPKEPTPTTI---K 173
           |    :||:|    |    ||                | |     ||||     |:    |
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy     174 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 227
           || :   ||   | ||| |   ||| :|  |       || |   ||| :|  ||   |
Db     126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy     228 SAPTTPKEPAPTTP----------KKPAPTTPKEP------APTTPKEPTPTTP------ 265
           ||| |   |  | |           :||| :   |      || |   |  | |
Db     180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy     266 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 306
               ||| :  |        || |   |  |||       || |   |  | |
Db     240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy     307 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 361
           || |   |  | |     |:   |: || :|  ||   | ||| |   |:| :|  |
Db     300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy     362 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 409
               || |   |  | |       || |   |  | |       || |   |  |||
Db     357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy     410 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 469
           |:  :| |       |       || |   |  | |   |       | |   |  |
Db     417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy     470 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 521
           | |    :||:|   |  | |       || |   |  | |   |     :|| |  |
Db     473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy     522 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 569
           | |           :|||   ||       ||    :|||  | |     |:    ||
Db     529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy     570 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 615
           |  | | |       :|| |   |  | |           :|||   ||    | ||
Db     584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy     616 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 659
             :|||  | |       :|| |   |  | |           :||| :  | ||    |
Db     643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700

Qy     660 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 712
           | |   | :       :  :||       |  |    : ::|      |   | :| |
Db     701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753

Qy     713 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 765
           |   |:  | |  | :|       |:||    :|       |       |:   |
Db     754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807

Qy     766 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 821
           | |    :|:   |  ||  |       |:  :    |  || |    |::  | |
Db     808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867

Qy     822 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 881
           |   : ||      |  | | |     | :|  |    :: |:  |              |
Db     868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA------------PG 909

Qy     882 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 940
           | | :  |   | :  | |:     |           ||   |  | : :      | :
Db     910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy     941 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 981
           : |:          | : |||           :||| ||
Db     956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996
```

```
RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
```

```
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;

  Query Match           10.3%;  Score 556.5;  DB 1;  Length 875;
  Best Local Similarity  28.7%;  Pred. No. 5.2e-16;
  Matches  272;  Conservative  106;  Mismatches  375;  Indels  195;  Gaps   56;

Qy      42 TQHNKVSTSPKITTAKPINP-----------RP--SLPPNSDTSKETSLTVNKETTVET 87
           |:|  |    | : : |  |            ||  | ||    ||   | : |: :
Db       1 TKHEPVYKPVKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYG-SKTNYLPLAKKLSSYK 59

Qy      88 KETTTTNKQTSTDG--KEKTT---SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKG 142
              || | :|:     | | |   : |   |   |          |:   || ||: :
Db      60 PIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYPPTYKSKPTYKPKITYPPTYKAKPSY-- 117

Qy     143 PALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTT 193
           |:   ||:  | | |     | :  ||   | | |  |: |     |:  |:: |: |:
Db     118 PSSYKPKKTYPPTYKPKLTYPPTYKPKPSYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSY 177

Qy     194 PKEPAPTTTKEPA--PTTPKEPA--------PTTTKEPA-PTTTKSAPTTPK--EPAPTT 240
           |    ||   :|:  ||   :|:         ||   :|  |:| |: |: |   :  ||
Db     178 P----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTY 233

Qy     241 PKKPA--PTTPKEPA--PTTPKEPT--PTTPKEP---APTTKEPAPTTPKEPA--PTAPK 289
             ||:  ||   :|:  ||   :|:  ||   :|   |  | :  ||   :|:  ||
Db     234 KAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYPPTYKA 293

Qy     290 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 339
           ||:  ||   :|:  ||   :|:  ||   :||  ||   :|: | | |: ||   :|
Db     294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy     340 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 395
           |:| |: |: |    ||   :|:     :  ||   ||:  ||   :|:  ||   :|:
Db     354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy     396 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 455
             | |  ||   :|       | |:|  |  |: |    |:   ||:     :  ||
Db     410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy     456 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 509
           :||        |:| ||     |  ||    |  | | |     |:   ||:| | |
Db     458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy     510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 565
           || |   | | | |: |    ||   | | |              || : | |    |
Db     507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy     566 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 617
           ||    |   | | |      ||   :|:  ||   ||:   |   |:| |  |:   :
Db     552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy     618 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 666
           |  |: |    ||   :|:  ||   ||  |:|    |  ||    ::| | | |
Db     612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy     667 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 715
           | | :|:     |   |:|| ||   :|     :|  | |    |:       ||    |
Db     668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy     716 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 775
           | | | :   | | |  | |  | :   |   | ::|        :   |    |
Db     727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy     776 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 832
           | |: :: |     |||   |    ||: |  |  :        |||  |    |   |
Db     778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826

Qy     833 TSTKKPKTMPRVRKPKTTPTP-----------RKMTSTMPELNPTSR 868
           :: |   : |   | ||:  |            :  ||  |   ||:|
Db     827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874


RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
```

```
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;
```

```
  Query Match             10.3%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity   32.6%;  Pred. No. 1.4e-15;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy      58 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 117
           |:   | ||| |  |      |   |||  |:  | : |             | | ||||
Db     445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy     118 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 172
           |    ||      |  |    |: | ::| |   || | |  :   ::||     || :
Db     488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy     173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 231
             || | |  ||     ||   |     |:|  |||| |   |  :||  ||   | ||
Db     540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy     232 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
             ::|:  | |   ||   |    : ::||    |   || |   || | | :  :|:|
Db     597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy     292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
           |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db     648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy     352 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 406
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| |   |
Db     702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy     407 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 456
           ::|  :| | | ||     :|:||| | |||       || | :    : |:|  | | |
Db     758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy     457 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 516
             :||:|  :| |           || | |  || | |       ||  :| | | ||
Db     815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy     517 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 570
           ::|  :  |   | |    :|::||  |       ::||   | | | ||  ::
Db     856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 628
           |  |:||   | |   :   ||:||       :  |||   :||    | |    |:
Db     912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTSRSSTGMSCP-PNARYESCAC 966

Qy     629 PTTPKEPAPT 638
           | : | | |:
Db     967 PASCKSPRPS 976
```

```
RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
```

```
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73        H-A-P-P.
FT   REPEAT      148    151        H-A-P-P.
FT   DOMAIN      229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235        1.
FT   REPEAT      236    242        2.
FT   DOMAIN      205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             10.2%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 6.6e-16;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy     136 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 184
           || ||: |  |  |   :| |::     |  | :| |     |:|   : ||     | |
Db      24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy     185 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 235
           :    |:    |       |:    | |:   :|  | |:    |        |:  :
Db      81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy     236 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 290
           |       |:|:    |     | |:   | : : |:| ::   | |.|  :| ||
Db     141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy     291 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 345
           |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db     194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy     346 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 399
              |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db     250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy     400 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 449
             | ||   :|  || : |   :|  |  |  |::      |   :|  |   :|  |  |
Db     310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy     450 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 506
           |  |  | | :   | ||  ::  |    :|  ||| |   | ||    | | |  :
Db     370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy     507 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 566
           |       | |   | |     | |    |  | |      :| |  |  : |     |
Db     428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy     567 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 626
           | :|    |       | |   | |     |   : || |    |:| |
Db     482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy     627 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
```

```
           |     | |    |: | ||  : | | |     |     |   |: |   ||
Db     539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy     685 PTPKALENSPKEPGVPTTKTPAATKP 710
           |:|           ||| :| :  |
Db     599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
```

```
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;


  Query Match             9.8%;  Score 533;  DB 1;  Length 1087;
  Best Local Similarity   27.5%;  Pred. No. 5.3e-15;
  Matches  216;  Conservative   81;  Mismatches  330;  Indels  158;  Gaps   38;

Qy      21 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 75
           :|   |      | :|      |  ||  | || :   |   :             : |||
Db     404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy      76 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 135
           :   :   :    |  |  :     |:|   : :  :  |:  |   :| ::  | |  :
Db     450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy     136 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 195
           ::  |  | :  :  :|   | ||    :  :|   || |   | |    : : | :|
Db     510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy     196 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 245
           || :|   | ||   || ||  | | |    :|:  || |   | ||    |   | ||
Db     565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy     246 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 290
              |   | ||    | | | |    :|:  | ||   :  :|   | ||    |   |
Db     620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy     291 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 342
           |    :|  || ||    | ||    :  ||   | ||    |   || |   | ||
Db     677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy     343 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 397
           : :|  :| |  ||   | ||    |   | |      : | |  || | :  :  :|
Db     730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy     398 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 457
           | |    | ||     | |  :| |     |||  |   :  |   || :   | | :
Db     790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy     458 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 505
           |  ||:  :|   || :|   || | |:    ||:    :|       |||     :|
Db     849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy     506 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 563
           |||  |  : |: ||    ||  :|   | |::|  :|::  |      |    :
Db     905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964

Qy     564 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 623
           : ||    || |   ||    ||  |   |  |:|:     || |: |
Db     965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP------------ 1006

Qy     624 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 683
                 ||  |  |:|                    | :| ||  : |:|
Db    1007 --------KKEEMPAAPEK-----------------KDTKEEKTTESRKPEEK------ 1034

Qy     684 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 743
           || :|   ||     :| |  :||:     |  :|::     ||      | |:: |
Db    1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy     744 TTTEK 748
           |  ||
Db    1083 TKGEK 1087


RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
```

```
CC   or send an email to license@isb-sib.ch).
CC   ------------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;


  Query Match             9.8%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 7e-15;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy     152 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 210
           ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db     599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy     211 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 257
           :  |  |   |  ::  | |:|| :  | :||  |       ||:|| : :    |:||
Db     652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy     258 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 316
           : :   ||  || ::    |:|| : :       | :||  || ::      |     :
Db     712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy     317 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 372
             ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db     765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy     373 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 427
           ||    | :||  |  ::      |     | : |  || ::   | |:||       ||
Db     818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy     428 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 477
           :|| :  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db     877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy     478 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 537
           |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db     936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

Qy     538 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 596
             || :|  | :|     |:|| :  | :||  || ::   |  | :   | :||
Db     984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy     597 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 652
           |     | ::  |  || :|  | :||      :| :||  || ::    |  || ||
Db    1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy     653 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
           :::  |||:|  : :: | :|  ||| : ||  ||                 ||| :
Db    1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy     712 MT 713
           |
Db    1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ------------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;


  Query Match             9.8%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 5.8e-15;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy     123 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 178
           :| |  :| | ||    |      ||   |       |  | : ||  | ||     | |
Db       8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy     179 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 233
```

```
Db     62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy    234 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 279

Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    280 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 331

Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    332 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 387

Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    388 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 428

Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    429 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 484

Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    485 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 533

Db    414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    534 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 585

Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    586 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 631

Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    632 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 681

Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 739

Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    740 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 792

Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    793 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 852

Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy    853 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 895

Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864
```

```
RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN       21     41       NONREPETIVE LINKER.
FT   DOMAIN      124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                                P-[PST]-[ST].
FT   DOMAIN      184    192       NONAPEPTIDE 1.
FT   DOMAIN      213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;
```

```
  Query Match             9.6%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity   29.2%;  Pred. No. 1.7e-14;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy     19 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 69

Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy     70 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 129

Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    130 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 188

Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    189 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 243

Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    244 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 296

Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    297 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 351

Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    352 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 401

Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    402 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 456
```

```
Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    457 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 507

Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    508 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 564

Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    565 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 620

Db    584 -------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    621 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 669

Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    670 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 725

Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    726 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 785

Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    786 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 840

Db    786 I-------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy    841 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 882

Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871


RESULT  12
MUC1_XENLA
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;
```

```
  Query Match             9.4%;  Score 509.5;  DB 1;  Length 662;
  Best Local Similarity   28.1%;  Pred. No. 3e-14;
  Matches  223;  Conservative   48;  Mismatches  276;  Indels  247;  Gaps   30;

Qy      92 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 151
           ||    :  ||: | :|: : : |||:|       :| | ||  |     |   ||
Db       3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy     152 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 211
                 |: |    |     ||||   || |    ||||    ||||    ||||
Db      52 --------AAATAAAETTAAAGEAPTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy     212 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 269
           || |   | |||  | ||||      :|    ||||    | :  :  : :: :|
Db     104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy     270 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 314
             |: :|   :   | |:       ||     ||          :|   : :|   ||||
Db     164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy     315 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 374
                 |||| : | |||    ||||            |||  :  |||
Db     219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy     375 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 434
                       ||  :  ||||              || | |  ||    ||||
Db     246 ------------TTTTKATPTTT------------TTTKATTTTT----TPTT----- 270

Qy     435 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 494
                   ||    | |||   | |||
Db     271 ---------TTTTTKATTTPTTTTTTTP------------------------------- 289

Qy     495 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 554
                     |  ||   |  ||       | |           :|   :    :
Db     290 ----------TTTTTKATTTTTTTSGECKMEPSK-----------REDCGYSGITESQCR 328

Qy     555 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 604
           ||     : | |               ||:     |    |    |    :|    : :
Db     329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy     605 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 664
            : ||    :||   |  |   | |||   |  ||  |   |||       |:  :|||||
Db     385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTPTTTTTTKATTTTP------TTTTTPTTT 436

Qy     665 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 724
           ||  .              | ||                || ||  | |  |||   | |
Db     437 TTTTT---------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy     725 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 784
           |||  |||   |     | |||: |       |||  |:|||  ||  |  ||   ||
Db     464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy     785 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 840
           ||||  | |||    | |    |  ||        : |  :|    :|  : |    :
Db     511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy     841 MPRVRKPKTTPTPR 854
           :|:|   |  |: |
Db     567 LPQVADCKVAPSSR 580


RESULT  13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;


  Query Match             9.3%;  Score 503.5;  DB 1;  Length 1970;
  Best Local Similarity   33.3%;  Pred. No. 1.2e-13;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
           |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
```

```
          |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
          :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
             :||:|   :||:    :||:|   :||:|   | ||:|   :||:|   :||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
          ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
          :|   :||:|| |          ||:|   |:  |  ||: |   | || | :|||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy    560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
          ||:|   :|||||  :||:|  |   |      ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
          :|||:|   :||:||    :||:|   :||:|      |  :::|           .|||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy    677 STPE 680
          |  |
Db   1966 SDEE 1969
```

```
RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;
```

```
  Query Match             9.2%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 2e-13;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;

Qy    151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
          |:|    || |   :  ||     :|:      |     :|:  :     |  :
Db   1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy    205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
          | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :|||:|   :|
Db   1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy    263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
          |:|   :||:   :||:|   :||:|   :||:|   :||:|           |||:|
Db   1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy    323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
          :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db   1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy    383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
             :||:|   :||:    :||:|   :||:|   | ||:|   :||:|   :||:|
Db   1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy    443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
          ||:|   :||:|   :||:|   :||:| : :|::|:     ||: |   ||  |  :|:
Db   1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy    501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
          :|   :||:|| |          ||:|   |:  |  |.: |   | || | :|||:||  :
Db   1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy    560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
          ||:|   :|||||  :||:|  |   |      ||:||       :||:   || | || |
Db   1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy    619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
          :|||:|   :||:||    :||:|   :||:|      |  :::|           .|||:
Db   1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965
```

```
Qy     677 STPE 680
           | |
Db    1966 SDEE 1969


RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;


  Query Match             9.2%;  Score 497;  DB 1;  Length 831;
  Best Local Similarity   27.6%;  Pred. No. 1.1e-13;
  Matches  195;  Conservative   59;  Mismatches  307;  Indels  146;  Gaps   35;

Qy      37 PDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQ 96
           |  ||  | || :  ||   :             : |:| :   :   ::  |  |  :
Db     197 PSMST--HIKVKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEED 242

Qy      97 TSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----T 152
             |:|:  : :  :    ||     | ::  | |    |  || |     | |   :
Db     243 KEAQGEEEEEAEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKS 293

Qy     153 PTTPKEPA-STTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTK 203
           |   | || : :| |     :  :|   | ||     |   || |   | ||    | | |
Db     294 PAEAKSPAEAKSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVK 352

Qy     204 EP----APTTPKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA-- 245
           |    :|   | ||     |   | ||     | : || |   | ||     |   | ||
Db     353 SPGEAKSPAEAKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATV 411

Qy     246 --PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTA 287
             |   | ||    | | | |    :|   | || : | |    :|   | ||     |
Db     412 KSPVEAKSPAEVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVV 470

Qy     288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEP 338
           | ||    | ||   |   | :  :  ||   | ||    |   ||     :|   | |
Db     471 AKSPAEA--KSPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSP 528

Qy     339 APTTTKS-----APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTT 389
             ||     :|   | |    | ||    |   | |    |  |     :
Db     529 VKEGAKSLAEAKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPE 588

Qy     390 PKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEEL 449
           | ||    |:||    |   :  || | :  |: | |  ||:||   |  :|   ||:
Db     589 AKPPAKEEAKRPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV 644

Qy     450 APTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTT 509
```

```
              :||     ||  | |||     : |: ||   ::|     |||    ||::
Db    645 ---KAKEPPKKVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEA 700

Qy    510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPT 569
          |  |    ||     |::  | :|      :|     |  |            |||:
Db    701 KKEEA----KEKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKP 744

Qy    570 TPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAP 629
          : ||     ||       ||  |  |:|   ||           ||  |     ||  |
Db    745 SEKE----KPK-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAK 784

Qy    630 TTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
            ::       |  |  |  |    :| |: |  |:     |:|::
Db    785 AKEED-------KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824
```

Search completed: April 26, 2002, 16:20:03
Job time: 388 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:17:20 ; Search time 49.78 Seconds
(without alignments)
457.932 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1212
Perfect score: 5416
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....SPIDTVFTRCNCEGKTFFFK 1013

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	9.6	805	4	US-09-103-429A-4	Sequence 4, Appli
2	508	9.4	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	9.4	744	6	5202236-25	Patent No. 5202236
4	498.5	9.2	786	4	US-09-103-429A-3	Sequence 3, Appli
5	489	9.0	1721	3	US-08-700-651-5	Sequence 5, Appli
6	489	9.0	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	9.0	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	9.0	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	8.8	652	6	5202236-13	Patent No. 5202236
10	452	8.3	960	4	US-09-219-849-5	Sequence 5, Appli
11	424.5	7.8	1867	2	US-08-479-537A-5	Sequence 5, Appli
12	424.5	7.8	1867	4	US-09-083-116-5	Sequence 5, Appli
13	424.5	7.8	2035	2	US-08-479-537A-2	Sequence 2, Appli
14	424.5	7.8	2035	4	US-09-083-116-2	Sequence 2, Appli
15	419.5	7.7	2476	2	US-08-276-967-2	Sequence 2, Appli
16	419	7.7	1185	4	US-09-041-886-23	Sequence 23, Appl
17	417	7.7	829	1	US-08-642-255-132	Sequence 132, App
18	417	7.7	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	7.7	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	7.7	837	1	US-08-477-509B-103	Sequence 103, App
21	417	7.7	837	1	US-08-642-255-101	Sequence 101, App
22	417	7.7	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	7.7	837	3	US-08-482-085B-103	Sequence 103, App
24	417	7.7	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	7.6	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	7.6	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	7.5	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	7.5	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	7.5	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	7.4	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	7.4	682	1	US-08-397-633A-36	Sequence 36, Appl
32	381	7.0	1848	4	US-08-296-791-6	Sequence 6, Appli
33	381	7.0	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
34	376	6.9	960	4	US-09-219-849-6	Sequence 6, Appli
35	374.5	6.9	1537	1	US-08-325-267A-2	Sequence 2, Appli
36	373	6.9	1231	3	US-08-904-263A-4	Sequence 4, Appli
37	371	6.9	761	2	US-08-707-237A-84	Sequence 84, Appl
38	371	6.9	762	1	US-08-642-255-114	Sequence 114, App
39	371	6.9	762	1	US-08-397-633A-26	Sequence 26, Appl
40	370.5	6.8	1064	1	US-08-642-255-62	Sequence 62, Appl
41	368	6.8	762	1	US-08-642-255-120	Sequence 120, App
42	368	6.8	762	1	US-08-397-633A-31	Sequence 31, Appl
43	367.5	6.8	1187	1	US-08-320-559-28	Sequence 28, Appl
44	367.5	6.8	1187	3	US-08-545-860D-28	Sequence 28, Appl
45	367.5	6.8	1187	5	PCT-US94-04496-28	Sequence 28, Appl

ALIGNMENTS

```
RESULT 1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-4

  Query Match             9.6%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity   27.8%;  Pred. No. 2.4e-26;
```

```
 Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy    125 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 179
          |: |    |:     | |||  |    |  | |     | |   |              |
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    180 KEPAPTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 227
          :: || |             :  | |  || ||    || | :  |||||:  |||||
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    228 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 287
           ||||  :   ||           |||| : || |     | ||: |  ||
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 347
              ||||  : ||||  : |||||:          |||||: |||||:  ||||   |
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    348 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 391
          |||   :|  |  || |||    || |||  ||||:    ||  |
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    392 ------------------------------------------------EPAPTTTK 399
                                                           |||  |:
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    400 K----------------PA---------------------------------------- 402
                             ||
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    403 --------------------------------PTAP--KEPAPT-----------TPKET 417
                                           || |   || |             | |:
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    418 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 443
                                   :| |    :| |              :    |     ||
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    444 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 503
          |     |||      |:|   || | |  |||  |    ||    ||||     ||||
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    504 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
            :|||   |   |||     ||||    |  |:  | |  ||:  :   |||     ||
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    561 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 619
          ||     |||       |:|   ||||   ||| |           | |  ||:  |:  |
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    620 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 679
          ||     |  | |   ||||   ||| |  | ||| :  : |     ||  |          |
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    680 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
            :| |            | ||  || | ||
Db    686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             9.4%;  Score 508;  DB 3;  Length 1837;
  Best Local Similarity   21.3%;  Pred. No. 4.7e-25;
  Matches  295;  Conservative  100;  Mismatches  494;  Indels  496;  Gaps   48;

Qy     33 KVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTT 92
          | ||  |:||     :|:   ||        :   : |:  |: |    ||  |  |||
Db    307 KHTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 366

Qy     93 TNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPT 152
          |   |:|  |: ||:   | :   |:      |:      |    |||     || |  |
Db    367 TTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 426

Qy    153 PTTPKEPASTTPKEPTPTTIKSAPTTPKEP--APTTTKSAPTTPKEPAPTTTKEPAPTTP 210
            ||     :||    | ||  :  || |:|    ||| :  ||      |||     ||
Db    427 TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 486

Qy    211 KEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE----------- 259
             | ||||:|  ||| :  || |:|  ||      || :  :   | |
Db    487 TTTATTTTKKPTTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECE 546

Qy    260 ------------------------PTPTT---------------------------PK 266
                                     |   |                          |
Db    547 AKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPD 606

Qy    267 EPAPTTKEP--------------------APTTPKEPA-PTAPKKP------------- 291
            |  |                       :|   |: :  :|| :|
Db    607 RSLDFTIPPVAGHNSCSIIVGVSGGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCT 666

Qy    292 ---------------APTTPKEPAPTT---PKEPAPTTTKEPS-PTTPKEPAPTTTKS- 330
                          ||  :|  ||    | :|  |||  || |||      ||| :
Db    667 AKYGAIHSGYQTSADFVTTTTAKPTTTTTGAPGQPTTTTTGSPSKPTTTTTTKATTTTTI 726

Qy    331 ----APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPA 386
                 |||::|  |||    |  |   : ||| :|  ||    | ||      |||
Db    727 LNPIITTTTQKPTTTTTTKVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT----- 781

Qy    387 PTTPKEPAPTTTKKPAP-------------------------------TAPKEPAP--- 411
            || |    |||  | |                                 :  || |
Db    782 TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQ 841
```

```
Qy    412 -------------TTPKETA-PTTP-------------------------------KKL 425
Db    842 AGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNL 901

Qy    426 T-----------------PTTPEKLAPTTPE---------------------------- 439
Db    902 TVDTDETYGLPIDTLTGYPLDPVSLIPFNPETGELFDPISDEIMNGTIAGIVSGISASES 961

Qy    440 ----KPAPTTPEE----------LAPTT---------PEEPTPTTP--EEPAPTTPKAAA 474
Db    962 LLSQKSAPIDPATNMVVGEFGGLLNPATGVMIPGSLGPSEQTPFSPEIEDGGIIPPEVAA 1021

Qy    475 PNTPKEPAPTTPKEPAPTTPKE---------------------------------PA 498
Db   1022 ANADKFKLSIPPSVPESIPEKDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNPI 1081

Qy    499 PTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGT 558
Db   1082 MKTPTQTDSVTGKPIDPTT---GLPFNP--PTGHLINPTNNNTMDSSFAGAYKYAVSNGI 1136

Qy    559 APTT----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----------KKPAPK-- 601
Db   1137 KTDNVYGLPVDEITGLPKDPVSDIPFNSTTGELVDPSTGKPINNYTAGIVSGKRGLPPIE 1196

Qy    602 ----ELAPTTTKGP----------TSTT-------SDKPAPTTPKETAPTTPKEPA--- 636
Db   1197 DENGNLFDPSTKLPIDGNNQLVNPETNSTVSGSTSGSTKPKPGIPVNGGGVVPDEEAKDQ 1256

Qy    637 ----------PTTPKKPAPTTPETPPPTTSEVSTPTTTKE-PTTIHKS---PDESTPEL 681
Db   1257 ADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNIINPETGKVIPGSLPGSLNYPSFNTPQQ 1316

Qy    682 SAEPTPKALE---NSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKM 738
Db   1317 TDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP-I 1375

Qy    739 TKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE- 797
Db   1376 DLETGLPRDP------VSGLPQLPNGTLVD----------------PSNKKPIPGSHS 1411

Qy    798 --IMNKPEETAKPKDRATNSKATTPKPQKP-TKAP--KKPTSTKKPKTMPRVRKPKT--- 849
Db   1412 GFINGTSGEQSHEKDPSTG------KPLDPNTGLPFDEDSGSLINPETGDKLQGSHSGTF 1465

Qy    850 TPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPK-SEDA---------- 898
Db   1466 MPVPGKPQGENGGIMTPEQILEAL-------NKLPTSNEVNISPRPSSDAVPDRPTNTWW 1518

Qy    899 ---------------GGAEGETPHMLL------RPHVFMPEVTPDMDYLPRVPNQGIII 936
Db   1519 NKISGQTFQVDGKKTIPGSAASVIHTALGTPTQTDPTTGLP--SDPSTGLPFIPGFNVLV 1576

Qy    937 NPMLSD-------------ETNIC-----------NGKPVDGLTTL---RNGTLV-AF 966
Db   1577 DPQTGEQIKGSVPYVSLYVKEKNIVTEAAYGLPVDPKTGFPIDPISYLPFAKNGELIDPI 1636

Qy    967 RGHYF 971
Db   1637 SGKYF 1641


RESULT   3
5202236-25
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
;      LENGTH: 744
5202236-25


  Query Match             9.4%;  Score 506.5;  DB 6;  Length 744;
  Best Local Similarity   29.9%;  Pred. No. 2.2e-25;
  Matches  239;  Conservative   78;  Mismatches  308;  Indels  175;  Gaps   50;

Qy     51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
Db     24 PKMTYPPTYKPKPSYPP----------------TYKSKPTY----------KPKIT---- 53

Qy    111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
Db     54 -----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 275
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    276 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 324
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    325 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 377
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    378 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 429
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    430 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 478
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    479 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 532
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    533 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 591
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    592 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 646
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    647 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 693
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    694 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 739
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719

Qy    740 KETATTTEKTTESKITATTT 759
Db    720 YKAKPTYPSTYKAKPTYPST 739


RESULT   4
US-09-103-429A-3
```

; Sequence 3, Application US/09103429A
; Patent No. 6187558
; GENERAL INFORMATION:
; APPLICANT: Granados, Robert R
; APPLICANT: Wang, Ping
; TITLE OF INVENTION: A No. 6187558el Invertebrate Intestinal Mucin
; TITLE OF INVENTION: cDNA and Related Products and Methods
; NUMBER OF SEQUENCES: 4
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Brown, Pinnisi & Michaels, P.C.
; STREET: 118 No. 6187558th Tioga
; CITY: Ithaca
; STATE: NY
; COUNTRY: USA
; ZIP: 14850
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/09/103,429A
; FILING DATE: 24-JUN-1998
; CLASSIFICATION: 800
; ATTORNEY/AGENT INFORMATION:
; NAME: Michaels, Christopher A
; REGISTRATION NUMBER: 34,390
; REFERENCE/DOCKET NUMBER: BTI-39
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (607) 256-2000
; TELEFAX: (607) 256-3628
; INFORMATION FOR SEQ ID NO: 3:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 786 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: protein
; HYPOTHETICAL: NO
; ANTI-SENSE: NO
; ORIGINAL SOURCE:
; ORGANISM: Trichoplusia ni
; TISSUE TYPE: peritrophic membrane
US-09-103-429A-3

Query Match 9.2%; Score 498.5; DB 4; Length 786;
Best Local Similarity 29.1%; Pred. No. 7.6e-25;
Matches 200; Conservative 25; Mismatches 207; Indels 255; Gaps 32;

```
Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 191
          | | || ||: || |      ||| : | :||          ||||  :   |||  ||
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    192 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 251
          || :  |||||:  | ||    ||||| : |||||  ||||  : |||| :  |||| :
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    252 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 308
           |||| : ||  |     |    || |||    | |   || |||  ||||: |  |
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    309 ----------------------------------------------PTTTKEPSPTTPK- 321
                                                         |   :  ||   :
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    322 ----EPAPTTT---------------------------------------------------- 328
              |||  |
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    329 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 361
                                      :|      | |  |    || | |  |     |
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    362 APTT-----PKEP----------------------APTTPKKPAPTT---------- 381
          |  :    | |                       :| |    :| |
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    382 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 438
              |||     |||     ||||   |:   |  | | |   : |||     ||
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    439 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 496
            ||||     ||||     :|||    | ||    ||| |  |: |   |    |||
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 556
           |||     ||||     |||:  |    |   ||   || ||             | |
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    557 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 616
            ||||   ||| |     ||     || |:                 || ||  | :|||
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    617 DKPAPTTPKETAPTTPKEPAPTTPKKP 643
           || |  |      |  | |  |  |
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699

RESULT 5
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
; APPLICANT: PETERSEN, CAROLYN
; APPLICANT: LEECH, JAMES
; APPLICANT: NELSON, RICHARD, C.
; APPLICANT: GUT, JIRI
; TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
; TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
; TITLE OF INVENTION: INFECTIONS
; FILE REFERENCE: 480.19-4(HV)
; CURRENT APPLICATION NUMBER: US/08/700,651B
; CURRENT FILING DATE: 1997-08-14
; EARLIER APPLICATION NUMBER: 08/415,751
; EARLIER FILING DATE: 1995-04-03
; NUMBER OF SEQ ID NOS: 15
; SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
; LENGTH: 1721
; TYPE: PRT
; ORGANISM: Cryptosporidium parvum
US-08-700-651-5

Query Match 9.0%; Score 489; DB 3; Length 1721;
Best Local Similarity 22.1%; Pred. No. 7.3e-24;
Matches 284; Conservative 82; Mismatches 476; Indels 442; Gaps 41;

```
Qy   5 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 64
          : | || | :| || | : | | | :| | : ||| :
Db   116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy   65 LPPNSDTSKE--------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
           | ||: | | | | || | |||| |:|
Db   159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTT---------- 208

Qy   110 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
                                      ||| || | || :|| |
Db   209 ---------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy   170 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 229
          || : || ||| : || ||| || ||| ||| :
Db   241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy   230 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 289
```
```

```
Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT---------- 350

Qy    290 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 349

Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTTKKPTTTTTATTTTT 406

Qy    350 PKE-------------------------------------------PSPTT------- 357

Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    358 ---------------------------TKEP--------------------APTTPKE 368

Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    369 PAPTTPKKPAPTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 410

Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    411 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 470

Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    471 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 530

Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTPL 680

Qy    531 PK----ELAPTTTKEPTS---------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 570

Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    571 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 627

Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    628 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 673

Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    674 ------------------------PDEST-------------PELSAEPTPKALENSPK 695

Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    696 EPGVPTTKTPAATKPEMTTTAKDKTTER----------------DLRTTPETTTAAPKM 738

Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    739 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797

Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    798 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 842

Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    843 --------------RVRKPKTTPTPRKMTS--TMPE---------------------- 862

Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    863 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 917

Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    918 EVTPD--------MDYLPRVPNQGIIINP-------------MLSDETNICNGK----- 950

Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy    951 ----------PVDGLTTLRNGTLV 964

Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281

RESULT   6
```

```
US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  6:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1721 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-6
```

Query Match 9.0%; Score 489; DB 3; Length 1721;
Best Local Similarity 22.1%; Pred. No. 7.3e-24;
Matches 284; Conservative 82; Mismatches 476; Indels 442; Gaps 41;

```
Qy      5 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 64

Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy     65 LPPNSDTSKE---------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109

Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTT---------- 208

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169

Db    209 ---------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    170 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 229

Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    230 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 289

Db    301 TTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTT---------- 350

Qy    290 KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTT 349

Db    351 ----TTTTTTTTTTTTTTTTTTTTTTTTTTTTKPTTTTTTTTTTTTTKKPTTTTTATTTT 406
```

```
Qy    350 PKE-------------------------------------------PSPTT------- 357
Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    358 ----------------------------TKEP--------------------APTTPKE 368
Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    369 PAPTTPKKPAPTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 410
Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    411 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 470
Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    471 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 530
Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPL 680

Qy    531 PK----ELAPTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 570
Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    571 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 627
Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    628 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 673
Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    674 -------------------------PDEST-------------PELSAEPTPKALENSPK 695
Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    696 EPGVPTTKTPAATKPEMTTTAKDKTTER-----------------DLRTTPETTTAAPKM 738
Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    739 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    798 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 842
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    843 ---------------RVRKPKTTPTPRKMTS--TMPE---------------------- 862
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    863 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 917
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    918 EVTPD--------MDYLPRVPNQGIIINP--------------MLSDETNICNGK----- 950
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy    951 ----------PVDGLTTLRNGTLV 964
Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281
```

RESULT 7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
; GENERAL INFORMATION:
; APPLICANT: Hoffman, Stephen L.
; APPLICANT: Charoenvit, Yupin
; APPLICANT: Hedstrom, Richard
; APPLICANT: Khusmith, Srisin
; APPLICANT: Rogers IV, William O.
; TITLE OF INVENTION: Protective malaria sporozoite surface protein
; TITLE OF INVENTION: immunogen and gene
; NUMBER OF SEQUENCES: 2
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: A. David Spevack
; STREET: NMRDC Building 1 T-12 National Naval
; STREET: Medical Center
; CITY: Bethesda
; STATE: MD
; COUNTRY: USA
; ZIP: 20814-5044
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.24
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/638,431
; FILING DATE: 19910110
; CLASSIFICATION: 424
; ATTORNEY/AGENT INFORMATION:
; NAME: Spevack, Avrom D.
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (301) 295-6759
; TELEFAX: (301) 295-4033
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 826 amino acids
; TYPE: AMINO ACID
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-07-638-431-2

Query Match 9.0%; Score 488.5; DB 1; Length 826;
Best Local Similarity 26.4%; Pred. No. 3.5e-24;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649
```

```
Qy    631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
          |  |           |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    685 PTPKALENSPK-------------------EPGVPTTK 703
          |     |  ||                   || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;     APPLICANT:  Hoffman, Stephen L.
;     APPLICANT:  Charoenvit, Yupin
;     APPLICANT:  Hedstrom, Richard
;     APPLICANT:  Khusmith, Srisin
;     APPLICANT:  Rogers IV, William O.
;     TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;     TITLE OF INVENTION:  immunogen and gene encoding
;     NUMBER OF SEQUENCES:  2
;     CORRESPONDENCE ADDRESS:
;       ADDRESSEE:  A. David Spevack
;       STREET:  NMRDC Building 1 T-12 National Naval
;       STREET:  Medical Center
;       CITY:  Bethesda
;       STATE:  MD
;       COUNTRY:  USA
;       ZIP:  20814-5044
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.24
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  PCT/US92/00018
;       FILING DATE:  19920103
;       CLASSIFICATION:  424
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Spevack, Avram D.
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (301) 295-6759
;       TELEFAX:  (301) 295-4033
;   INFORMATION FOR SEQ ID NO:  2:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  826 amino acids
;       TYPE:  AMINO ACID
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  protein
PCT-US92-00018-2


  Query Match             9.0%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 3.5e-24;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
          ||  ::  |  |  |  |      |: |:||  |    :|     |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
          |      :  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | :|:     |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
            ||   |  ||  || |  |     |: | :| |       |: | :| |     : |:
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
            :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
```

```
          |:       |  |  |  |: | |  |: |||                    |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
          : |  |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
                      |: |::| |:|  |:  |||    |:   |  |:   | |  |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
          ||||               |  : ||:|    : |   : |     :   |       |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
          |  |           |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    685 PTPKALENSPK-------------------EPGVPTTK 703
          |     |  ||                   || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   9
5202236-13
;Patent No. 5202236
;     APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;     TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;     NUMBER OF SEQUENCES: 39
;     CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;       FILING DATE: 25-MAY-1990
;       APPLICATION NUMBER: 82,456
;       FILING DATE: 07-AUG-1987
;       APPLICATION NUMBER: 933,945
;       FILING DATE: 24-NOV-1986
;       APPLICATION NUMBER: 650,128
;       FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;       LENGTH: 652
5202236-13


  Query Match             8.8%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 1.6e-23;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy     51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
          ||:|      |:|| ||                  | ::| |           | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
                     || |  |||:          |    ||:  | | |     | :  ||
Db     41 -----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy    167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
            | | || ||  ||   | | |: |: |    |  | |    ||  | | |  | |
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 278
            | |  |: |  |:  | |   || ||    | |      | :  | | :|  |
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196

Qy    279 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 329
            |    ||   ||:  ||   :|:  ||   :|:   | :  |: |      |   | | |
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    330 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 376
          : ||   :|        |:| |: || |    ||   :|:  ||   :|      ||   |
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308
```

```
Qy    377 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 432
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    433 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 487
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    488 EPA-PTT---PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 540
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    541 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 598
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    599 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 651
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    652 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    712 MTTTAKDKTTERDLRTTPETTTAAP 736
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645
```

RESULT 10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
; APPLICANT: VAN HEERDE, GEORGE V.
; APPLICANT: VAN RIJN, ALEXIS C.
; APPLICANT: BOUWSTRA, JAN B.
; APPLICANT: DE WOLF, FREDERIK A.
; APPLICANT: MOOBROEK, ANDREAS
; APPLICANT: WERTEN, MARC W.T.
; APPLICANT: WIND, RICHELE D.
; APPLICANT: VAN DEN BOSCH, TANJA J.
; TITLE OF INVENTION: SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
; TITLE OF INVENTION: SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
; TITLE OF INVENTION: PREPARATION THEREOF
; FILE REFERENCE: 2728-2
; CURRENT APPLICATION NUMBER: US/09/219,849
; CURRENT FILING DATE: 1998-12-23
; NUMBER OF SEQ ID NOS: 50
; SOFTWARE: PatentIn Ver. 2.1
; SEQ ID NO 5
; LENGTH: 960
; TYPE: PRT
; ORGANISM: Artificial Sequence
; FEATURE:
; OTHER INFORMATION: Description of Artificial Sequence: Illustrative
; OTHER INFORMATION: amino acid sequence
US-09-219-849-5

Query Match 8.3%; Score 452; DB 4; Length 960;
Best Local Similarity 25.6%; Pred. No. 9.3e-22;
Matches 173; Conservative 60; Mismatches 309; Indels 134; Gaps 41;

```
Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 189
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    190 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 245
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    246 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA---------------- 284
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    285 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 330
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    331 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 387
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    388 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 437
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    438 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 492
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    493 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 535
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    536 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 590
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    591 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 642
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    643 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 693
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    694 PKEPGVPTTKTPAATK 709
Db    696 SRDPGPPGAHGPAGPK 711
```

RESULT 11
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
; GENERAL INFORMATION:
; APPLICANT: CHAMBON, Pierre
; APPLICANT: KIENY, Marie-Paule
; APPLICANT: LATHE, Richard
; APPLICANT: HAREUVENI, Mara
; TITLE OF INVENTION: PHARMACEUTICAL COMPOSITION FOR THE
; TITLE OF INVENTION: TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
; NUMBER OF SEQUENCES: 5
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: BURNS, DOANE, SWECKER & MATHIS, L.L.P.
; STREET: P.O. Box 1404
; CITY: Alexandria
; STATE: Virginia
; COUNTRY: United States
; ZIP: 22313-1404
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/479,537A
; FILING DATE: 07-JUN-1995
; CLASSIFICATION: 514
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: FR 90/13101
; FILING DATE: 23-OCT-1990
; PRIOR APPLICATION DATA:

```
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             7.8%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 1.1e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
               :||| :|       |||      |           ||      |  | |      |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|       |||      |  | |       ||      |  | |      |:

Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
           :  | | :  |         ||      |  |||        ||      |  | |    |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
           ||      |  | |       ||      |  | |       ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |     | |      |:    ||      |  | |      |:  :  |     | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |     :||:    |  | |      |:  :  |  |  |||      | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||      ||    :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
           .|  | |       :||      |  | |          :| |   ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :|  |  | |      :||      |  | |          :| | :  | |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
               |  :       :  :||      |  |    : ::|      | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
               |       |  ||      :  :    |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
             : ||      |  |   |       | |    :||  |:  |                | |
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy     884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
            | :  |    |      |:     |           ||   |  |   :      |
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy     943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
           |:  : :|  | |    :|  |           |   | |        |: |  |
Db     958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  12
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
```

```
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5
```

```
 Query Match             7.8%;  Score 424.5;  DB 4;  Length 1867;
 Best Local Similarity   23.7%;  Pred. No. 1.1e-19;
 Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |     ||||    |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
           :||| :|         |||      |             ||     |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|        |||      | | |       ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |        ||     | |||       ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||      |||     | |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
           ||      | | |       ||      | | |       ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  |    | |      |:    ||      | | |     |: :  |    | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |    :||:    |  | |      |:  :  |  |  |||     | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |          :| |   ||      ||    :| |  | |     |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
           | | |      :||    | | |                :| |   ||     | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      | | |            :| | :  | |     | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :      :  :||      |  |     : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |: | | | :|  ||      |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                 |     |  | ||       :  :    |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
           : ||      |  |   |      | |    :||  |:  |                 | |
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy     884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
           | :  |    |       |:     |           ||   | |   :      |
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy     943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
```

```
          |:  : :|  | |    :|   |          |  | |        |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  13
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2


  Query Match             7.8%;  Score 424.5;  DB 2;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 1.2e-19;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||               | |     ||||      |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
             :||| :|         |||      |             ||      |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|        |||      |  | |          ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |         ||      |  |||        ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||      |||      | |          |||      | |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||      | | |        ||      | | |        ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |      |:    ||      |  | |      |:  :  |      | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
            | |     :||:    |  | |      |:  :  |  |  |||      | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||      ||     :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             | | |       :||      |  | |           :| |   ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762
```

```
Qy    708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy    943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

```
RESULT  14
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/083,116
;      FILING DATE:
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  08/479,537
;      FILING DATE:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repea
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CC
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-2
```

Query Match 7.8%; Score 424.5; DB 4; Length 2035;
Best Local Similarity 23.7%; Pred. No. 1.2e-19;
Matches 255; Conservative 69; Mismatches 511; Indels 241; Gaps 53;

```
Qy     75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
Db     16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy    134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
Db     66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy    177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
Db    123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy    218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
Db    181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy    265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
Db    241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy    312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
Db    301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy    361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
Db    360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy    413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
Db    420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy    469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
```

```
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy    764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy    943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

RESULT 15
US-08-276-967-2
; Sequence 2, Application US/08276967
; Patent No. 5851817
; GENERAL INFORMATION:
; APPLICANT: Hardy, Daniel M.
; APPLICANT: Garbers, David L.
; TITLE OF INVENTION: Species-Specific Egg-Binding Proteins of
; TITLE OF INVENTION: Sperm
; NUMBER OF SEQUENCES: 8
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Arnold, White & Durkee
; STREET: P. O. Box 4433
; CITY: Houston
; STATE: Texas
; COUNTRY: USA
; ZIP: 77210-4433
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.25
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/276,967
; FILING DATE: Submitted Herewith
; CLASSIFICATION: 530
; ATTORNEY/AGENT INFORMATION:
; NAME: Kitchell, Barbara S.
; REGISTRATION NUMBER: 33,928
; REFERENCE/DOCKET NUMBER: UTSD:418\KIT
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 713-787-1400
; TELEFAX: 713-789-2679
; TELEX: 79-0924
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 2476 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-276-967-2

Query Match 7.7%; Score 419.5; DB 2; Length 2476;
Best Local Similarity 32.5%; Pred. No. 3.2e-19;
Matches 163; Conservative 43; Mismatches 177; Indels 119; Gaps 31;

```
Qy    406 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 465
Db    313 PSETSVSTEKPVAPTE----KPTVPSEIYTIPTEKPMVHMEKPIVHT--EKPTVPT-EKP 365

Qy    466 APTTPKAAAPNTPKEPA----PTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 521
Db    366 TIPTEKSTVPT--KKPTVFKEPTLPPE-GPTVPAE-RPTTPPEGPAVPPKG--PTVLTE- 418

Qy    522 APTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPT-TPKEPAPTTPKEPAPTTPK 580
Db    419 ------------WPTSHTEKSTVHTEKPILPTGKSTIPTEKPMVPTKRT------TTP- 458

Qy    581 GTAPTTLKEPAPTTPKKPAPKELAPT-TTKGPTSTT---SDKPAPTTPKETAPTTPKEPA 636
Db    459 -TERTTIPAEKPTV---PIEKPMVPTERTTIPTERTTIPTEKPTVPTEKLTVPT--EKPI 512

Qy    637 PTTPKKPAPTTPETPPP-----TTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPT 686
Db    513 VPTEKPIVPTEKHTIPTEKLTVLTERTTTPTERTTIPTEKPTVPTEKPSVPT-EKPTVPT 571

Qy    687 PKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTT 746
Db    572 --------EEPTIPTEKLTVPT--ERTTTPTKRTTTPTIRTTTPTI--------RTTTPT 613

Qy    747 EKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETA 806
Db    614 ERTT------TPTIRTTTPTERTT---IPTKKTT------VPTEKTIIPT------ERTI 652

Qy    807 KPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT 866
Db    653 AP---------TTPQP-SPTLVPTQPAAVVMPST------SATTVTPRTTIASCP---PN 693

Qy    867 SRIAEAMLQTTTRPNQTPNSKL 888
Db    694 AHFERCACPVSCQ-SPTPNCEL 714
```

Search completed: April 26, 2002, 16:17:34
Job time: 245 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:18:59 ; Search time 82.85 Seconds
(without alignments)
978.271 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1263
Perfect score: 5698
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....VAALSTAKYKNWPESVYFFK 1064

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 219241 seqs, 76174552 residues

Total number of hits satisfying chosen parameters: 219241

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : PIR_68:*
1: pir1:*
2: pir2:*
3: pir3:*
4: pir4:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1062.5	18.6	3020	2	A43932	mucin 2 precursor,
2	950	16.7	1664	2	T18262	S-layer protein -
3	853	15.0	1489	2	T31108	cyst germination s
4	802	14.1	1274	2	T16251	hypothetical prote
5	789.5	13.9	1367	1	S48478	glucan 1,4-alpha-g
6	784	13.8	1188	2	S49915	extensin-like prot
7	776	13.6	2187	2	T30826	nascent polypeptid
8	672	11.8	1344	1	A35175	mucin 1 precursor,
9	671	11.8	1151	2	T18535	high molecular mas
10	659.5	11.6	1229	2	T25697	hypothetical prote
11	633	11.1	489	2	T11622	extensin class 1 p
12	632.5	11.1	7962	2	I38346	elastic titin - hu
13	632	11.1	761	2	C84672	hypothetical prote
14	629	11.0	3507	2	T34513	hypothetical prote
15	626.5	11.0	990	2	I51618	nucleolar phosphop
16	625	11.0	6642	2	T29757	protein UNC-89 - C
17	622.5	10.9	839	2	F75518	hypothetical prote
18	622	10.9	3570	2	T45025	mucin MUC5B, trach
19	620.5	10.9	971	2	T19431	hypothetical prote
20	607.5	10.7	801	2	T29018	hypothetical prote
21	607.5	10.7	924	2	S27923	gene LF3 protein -
22	605	10.6	379	2	S50125	larval glue protei
23	559.5	9.8	350	2	S22456	hydroxyproline-ric
24	556.5	9.8	856	2	T16543	hypothetical prote
25	556.5	9.8	875	2	S23760	polyphenolic adhes
26	555.5	9.7	1630	2	A53577	ascites sialoglyco
27	553	9.7	2232	2	T34434	hypothetical prote
28	551	9.7	620	2	S06733	hydroxyproline-ric
29	545	9.6	873	2	A47283	calphotin - fruit
30	542	9.5	369	2	S20500	hydroxyproline-ric
31	538.5	9.5	416	2	JU0465	extensin precursor
32	533	9.4	1087	1	QFMSH	neurofilament trip
33	532	9.3	756	2	T27642	hypothetical prote
34	530.5	9.3	1162	2	JH0557	exo-alpha-sialidas
35	530	9.3	865	2	A47282	calcium-binding pr
36	522	9.2	328	2	JQ0985	hydroxyproline-ric
37	522	9.2	1459	2	T32271	hypothetical prote
38	518	9.1	813	2	S70795	vsaA protein precu
39	518	9.1	866	2	T45462	membrane glycoprot
40	518	9.1	1072	1	A37221	neurofilament trip
41	512.5	9.0	867	2	T45463	membrane glycoprot
42	509.5	8.9	662	2	A45155	mucin FIM-C.1 - Af
43	508	8.9	1832	2	T31113	mucin-like glycopr
44	505	8.9	700	2	A54641	interspersed repea
45	504.5	8.9	606	2	A43427	neurofilament trip

ALIGNMENTS

RESULT 1
A43932
mucin 2 precursor, intestinal - human (fragments)
N;Alternate names: mucin SMUC-41
C;Species: Homo sapiens (man)
C;Date: 10-Mar-1993 #sequence_revision 12-Apr-1996 #text_change 05-Nov-1999
C;Accession: A49963; A45106; B45106; A43932; B33532; A61257; PQ0328; PQ0329
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Siddiki, B.; Kim, Y.S.
J. Biol. Chem. 269, 2440-2446, 1994
A;Title: Molecular cloning of human intestinal mucin (MUC2) cDNA. Identification of t
A;Reference number: A49963; MUID:94132002
A;Accession: A49963
A;Molecule type: mRNA
A;Residues: 1-639 <GU1>
A;Cross-references: GB:L21998
R;Gum Jr., J.R.; Hicks, J.W.; Toribara, N.W.; Rothe, E.M.; Lagace, R.E.; Kim, Y.S.
J. Biol. Chem. 267, 21375-21383, 1992
A;Title: The human MUC2 intestinal mucin has cysteine-rich subdomains located both up
A;Reference number: A45106; MUID:93016075
A;Accession: A45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 626-1895 <GU2>
A;Cross-references: GB:M94131; NID:g186395; PIDN:AAA59163.1; PID:g186396
A;Note: sequence extracted from NCBI backbone (NCBIP:116706)
A;Accession: B45106
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 2037-3020 <GU3>
A;Cross-references: GB:M94132; NID:g186397; PIDN:AAA59164.1; PID:g186398
A;Experimental source: colon
A;Note: sequence extracted from NCBI backbone (NCBIP:116698)
R;Toribara, N.W.; Gum Jr., J.R.; Culhane, P.J.; Lagace, R.E.; Hicks, J.W.; Petersen,
J. Clin. Invest. 88, 1005-1013, 1991
A;Title: MUC-2 human small intestinal mucin gene structure. Repeated arrays and polym
A;Reference number: A43932; MUID:91358717
A;Accession: A43932
A;Molecule type: DNA
A;Residues: 1343-1350,'L',1352-1411,'S',1413-1448,'P',1450-1503,'T',1505-1915 <TOR>
A;Cross-references: GB:M74027; NID:g188863; PIDN:AAA59875.1; PID:g188864
A;Note: sequence inconsistent with the nucleotide translation
A;Note: sequence extracted from NCBI backbone (NCBIN:55749, NCBIP:55750)
R;Gum, J.R.; Byrd, J.C.; Hicks, J.W.; Toribara, N.W.; Lamport, D.T.A.; Kim, Y.S.
J. Biol. Chem. 264, 6480-6487, 1989
A;Title: Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evi
A;Reference number: A33532; MUID:89197956
A;Accession: B33532
A;Molecule type: mRNA
A;Residues: 1916-2193 <GU4>
A;Cross-references: GB:M22405; NID:g188873; PIDN:AAA36334.1; PID:g188874
A;Experimental source: intestine
R;Jany, B.H.; Gallup, M.W.; Yan, P.S.; Gum, J.R.; Kim, Y.S.; Basbaum, C.B.

J. Clin. Invest. 87, 77-82, 1991
A;Title: Human bronchus and intestine express the same mucin gene.
A;Reference number: A61257; MUID:91086481
A;Accession: A61257
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 'T',1925-1948,'TTS',1952-1954 <JAN>
A;Experimental source: bronchus
R;Xu, G.; Huan, L.; Khatri, I.; Sajjan, U.S.; McCool, D.; Wang, D.; Jones, C.; Forstner,
Biochem. Biophys. Res. Commun. 183, 821-828, 1992
A;Title: Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-t
A;Reference number: PQ0328; MUID:92198477
A;Accession: PQ0328
A;Molecule type: mRNA
A;Residues: 2328-2468 <XUG>
A;Cross-references: GB:M86523
A;Experimental source: small intestine
A;Accession: PQ0329
A;Molecule type: protein
A;Residues: 2328-2342,'K',2344-2354 <XUG1>
C;Genetics:
A;Gene: GDB:MUC2
A;Cross-references: GDB:120203; OMIM:158370
A;Map position: 11p15.5-11p15.5
C;Superfamily: von Willebrand factor; von Willebrand factor type A repeat homology; von
C;Keywords: glycoprotein; intestine; tandem repeat
F;2766-2834/Domain: von Willebrand factor type C repeat homology <VWC>

Query Match 18.6%; Score 1062.5; DB 2; Length 3020;
Best Local Similarity 29.2%; Pred. No. 9.2e-39;
Matches 342; Conservative 74; Mismatches 442; Indels 315; Gaps 39;

```
Qy      32 FKVTTPDTSTTQH---NKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
           :::  |: :  :|     ::| |   |       |: |        || |    ||  |
Db    1240 WEICGPNGTVEKHFNICSITTRPSTLTTFTTITLPTTP--------TSFTTTTTTTTPTS 1291

Qy      89 ET--TTTNK--------------QTSTDGKEKTTS--AKETQSIEKTSAKD-------- 121
           |  :|| |                | ||  :         : ||  | ||
Db    1292 STVLSTTPKLCCLWSDWINEDHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQH 1351

Qy     122 ------------------------------------------LAPTSKVLAKPTPKAE 137
                                                       :   |     |:|
Db    1352 GQKVQCDVSVGFICKNEDQFGNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPT 1411

Qy     138 TTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-TTPKEPAPTTTKSAPTTPKE 196
           |||  |  |||  || ||   | :|||  |  ||    | |||  |  |||   |||
Db    1412 TTTTLPPTTTPSPPTTTTTTPPPTTTPSPPITTTTTPLPTTTPSPPISTTTTPPPTTTPS 1471

Qy     197 PAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAP-TTPKEPAPTTPKKPAPTTPKEPAPT 255
           | ||||  |  |||  |  |||  | |||| | | |||  |  :|   |  |||  |  |
Db    1472 P-PTTTPSPPTTTPSPPT-TTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTT 1529

Qy     256 TPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEP 315
           |   |  |||  |  |   | ||:     ||    | |||   | |||   |  |||  |
Db    1530 TTTPPPTTTPSPPTTTPITP-PTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSP 1588

Qy     316 SPTTPKEPAPTTTKSAP-TTTKEPAPTTTKSAP-TTPKEPSPTTTKEPAPTTPKEPAPTT 373
              |  | |||| | | |||  | |||| | | |||  |  :||  |  |||  |  ||
Db    1589 PTITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTT 1648

Qy     374 PKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKL 433
           |  |||  |  |||  |  |||  |  | |  |  |||  : |||    | |||
Db    1649 TTPPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTP--SPPTT----TMTTPS-- 1700

Qy     434 APTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT 493
             |||  |  ||    : |||  | |||   |:|||  :    |     |||   |  ||
Db    1701 PTTTPSSPITTTTTPSSTTTP-SPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTT 1759

Qy     494 PKEPAPTTP-----KETAPTTP-----------KG---------------------- 512
           |  |: | |       | ||||            |
Db    1760 PLPPSITPPTFSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAA 1819
```

```
Qy     513 ------------------------------------------------------------ 512

Db    1820 NISCRATMYPDVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECV 1879

Qy     513 TAPTTLKEPAPTTPKKPAPKELAPTT--TKEPTSTTSDKP----APT-TPKGTAPTTPKE 565
           | |||:     || : | |  :  ||  |  || |::  |    ||| ||  |  |
Db    1880 TQPTTM---TTTTTENPTPTPITTTTTVTPTPTPTSTQSPNGLQAPTPTPISTTTTVTPT 1936

Qy     566 PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKPAPTTP 624
           | ||  : |  |||  |  |    | ||  : |    :  |||  || | || |    ||
Db    1937 PTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTVLITTTTTMTPTPTPTSTKSTTVTP 1995

Qy     625 KETAPTTPKEPAPTTPKKP-------APTTPETPPPTTSEVSTPTTTK-----EPTTIHK 672
             |  |    | ||  : |          |   || |||     || |      | || :
Db    1996 ITTTTTVTATPTPTGTQTPTMIPISTTTTVTPTPTPTTGSTGPPTHTSTAPIAELTTSNP 2055

Qy     673 SPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETT 732
             |: |||: |   :    |::     :|      :| |  | ||   |:     : | :|
Db    2056 PPESSTPQTSRSTSSPLTESTTLLSTLPPAIEMTSTAPPSTPTAPTTTSGGHTLSPPPST 2115

Qy     733 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITT---LKTTTLAPKVTTT 789
           | :|  |    |||   : |  | :| | |:|:    ||  ::|   ::|| | |  ::
Db    2116 TTSPPGTPTRGTTT--GSSSAPTPSTVQTTTTSAWTPTPTPLSTPSIIRTTGLRPYPSSV 2173

Qy     790 KKTITTTEIMNKPEE----------------------------TAKPKDRATNSKATTP 820
                  :   | |                              : |    | || :||
Db    2174 LICCVLNDTYYAPGEEVYNGTYGDTCYFVNCSLSCTLEFYNWSCPSTPSPTPTPSK-STP 2232

Qy     821 KPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP 880
           | ||:  | |||   ||   |    |:             : | |  : :  :  | :
Db    2233 TPSKPSSTPSKPTPGTKPPECPDFDPPR------------QENETWWLCDCFM-ATCKY 2278

Qy     881 NQTPNSKLVEVNP------------KSEDAGG-----------AEGETPHMLL----- 910
           | |    ||  |              : ||  |              |: || :
Db    2279 NNTVEIVKVECEPPPMPTCSNGLQPVRVEDPDGCCWHWECDCYCTGWGD-PHYVTFDGLY 2337

Qy     911 ------RPHVFMPEVTPDMDYLPRVPNQGIIIN 937
                 :| : |::| :|      | |: |:
Db    2338 YSYQGNCTYVLVEEISPSVD------NFGVYID 2364
```

RESULT 2
T18262
S-layer protein - Clostridium thermocellum
C;Species: Clostridium thermocellum
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18262
R;Fujino, T.; Beguin, P.; Aubert, J.P.
J. Bacteriol. 175, 1891-1899, 1993
A;Title: Organization of a Clostridium thermocellum gene cluster encoding the cellulo
e.
A;Reference number: Z18847; MUID:93209931
A;Accession: T18262
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1664 <FUJ>
A;Cross-references: EMBL:X67506; NID:g296879; PID:g296881; PIDN:CAA47841.1

Query Match 16.7%; Score 950; DB 2; Length 1664;
Best Local Similarity 31.0%; Pred. No. 3.8e-34;
Matches 312; Conservative 104; Mismatches 361; Indels 230; Gaps 55;

```
Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |   ||||:          ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :  ||:|| || |    ||   ||     || : || |:   :  ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
             ||   || |:   ||||:   || | : || |: ||:|| ||      ||   || |
```

```
Db    865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy    302 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 354
          :    |:  || : ||:|: ||:|| || | |  || : || |:    |    ||   ||:
Db    920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy    355 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 409
              || : || |: ||:|| ||      ||   || |:    |:   |    ||   ||
Db    980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy    410 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 462
           |:   || |    ||::  ||      ||  ::| |:   || :  ||  :||||: ||
Db   1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy    463 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 519
          ||| ||   :  |    || |:    |:  ||  || |: ||:|  ||      ||
Db   1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy    520 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 578
          || |:   :| | :  || : ||| | | ::| ||      ||   || |:  ||:
Db   1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy    579 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 634
            ||   || |: ||::|            || | ||:| |:   || :  ||   |
Db   1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy    635 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 693
          | |:   :| |: ||| | || :    || : |||       || ||  | ||||
Db   1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy    694 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 751
            ||      ||: |  :|  | |  |: |  |     | :  |  : |   : | |
Db   1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy    752 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 811
          : | :    . | | :: ||       |||  |  :||                   |
Db   1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy    812 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 856
          :  |  :           |  ||  |  ||||| |     : :|  :  |  :
Db   1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453

Qy    857 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 908
                |  || | |   |  :           :||:     :::: ||:  :|:
Db   1454 YLRGYPDGSFRPERNITRAEAAVIF----------AKLL----GADESYGAQSASPYSD 1498

Qy    909 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 934
          |   |                    | |:         |  : :| :|  | |
Db   1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy    935 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 981
              | ||   |    | |    : :: ||:|       :||                | | :
Db   1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy    982 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1026
            :|   |    : |:    |    |    | |    || |  || |
Db   1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642


RESULT   3
T31108
cyst germination specific acidic repeat protein precursor - Phytophthora infestans
C;Species: Phytophthora infestans (potato late blight agent)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 22-Oct-1999
C;Accession: T31108
R;Goernhardt, B.
submitted to the EMBL Data Library, April 1998
A;Reference number: Z20986
A;Accession: T31108
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1489 <GOE>
A;Cross-references: EMBL:AF061185; NID:g3851513; PID:g3851514; PIDN:AAC72308.1

C;Genetics:
A;Gene: car90


  Query Match             15.0%;  Score 853;  DB 2;  Length 1489;
  Best Local Similarity   32.3%;  Pred. No. 5.3e-30;
  Matches  355;  Conservative   46;  Mismatches  475;  Indels  222;  Gaps   49;

Qy      9 TKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTT------QHNKVSTSPKITTAKPINPR 62
          |   || ||  |:|      :   :  |  :|:          | :     ::|   |
Db    338 TPYAPTEKPYDVEETTYVTEESTYAPTKSETNAPTERMHYAHIEKPCDTEVTMYAPTEET 397

Qy     63 PSLP---------------PNSDT----SKETSLTVNKETTVETKETTT---TNKQTSTD 100
            |                | :|    ::||: |  :|||     | ||    | | |
Db    398 TYAPTEETTYAPTEETTYAPTEETPYEPTEETTYTPTEETTYAPTEETTYAPTEKTTYAP 457

Qy    101 GKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPA 160
          :| | :  |    | |      ||| :    ||   | ||       |   ||  |   ||
Db    458 TEETTYAPTEETPYEPTEETTYAPTKETTYAPT---EETTYASTEETTYAPTEETTYAPA 514

Qy    161 STTPKEPTPTTIKSAPTTPKEPAPT-TTKSAPTTPKEPAPT--TTKEPAPTTPKEPAPTT 217
            || |||  |   |||      |||  |  |||      |||  ||  ||  || ||   |
Db    515 EETPYEPTEET-TYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEET 573

Qy    218 TKEPAPTTTKSAPT----------TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKE 267
          |  |   ||  |||           |   |   |  ||  || ||   |   ||  |
Db    574 TYAPTEETT-YAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYA 632

Qy    268 P--------------AP---TTKEPAPTTPKEP------APTAPKKPAPTTPKEPAPTTP 304
          |              ||   ||  ||  || ||       |||     |||     |||
Db    633 PTEETTYASTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEE 692

Qy    305 KEPAPT--TTKEPSPTTPKEP------APT-TTKSAPT----------TTKEP------A 339
            |||  ||  |:  || ||       |||  |  |||           ||  |      |
Db    693 TTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTYA 752

Qy    340 PT-TTKSAPT--TPKEPSPTTTKEP------APTTPKEPAPTTPKKPAPTTPKEPAPT-- 388
          ||  |  |||  ||  |:  || ||       |||     |||     |||     |||
Db    753 PTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEE 812

Qy    389 TPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEE 448
          || ||   ||  |    | ||   |    ||     ||     |||     |||
Db    813 TPYEPTEETTYAPTEETPYEPTEET--TYTPTEETTYAPTEETTYAPTEKTTYAPTEETT 870

Qy    449 LAPT--TPEEPTPTTPEEP------APT----------TPKAAAPNTPKEPAPTTPKEP- 489
          |||  || |||  |   |       |||           |  |     |   ||  || ||
Db    871 YAPTEETPYEPTEETTYAPTKETTYAPTEETTYASTEETTYAPTEETTYAPAEETPYEPT 930

Qy    490 -----APTTPKEPAPTTPKETAPTTPKGTAPT--TLKEPAPTTPKKPAPK-ELAPT--TT 539
               |||     |||     |||     |||  |   ||  || :|  :   |||  ||
Db    931 EETTYAPTEETTYAPTEETTYAPTEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETT 990

Qy    540 KEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEP------APTTPKGTAPT-------- 585
            ||  |   |   |         |   ||  || ||       |||     |||
Db    991 YAPTEETMYAPIEETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAST 1050

Qy    586 ----------TLKEPAPTTPKKP------APKE---LAPT--TTKGPTSTTSDKPAPTTP 624
                    |   ||  || :|      || |    |||  ||  ||  |:  |   |
Db   1051 EETTYAPTEETTYAPAEETPYEPTEETTYAPTEETTYAPTEETTYAPTEETTYAPTEETT 1110

Qy    625 KETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT--TIHKSPDESTPELS 682
            |  || ||   |   |   |   |   |        ||  ||  | : :| |:|
Db   1111 YAPAEETPYEPTEETTYAPTEETTYAPTEETMYAPIEETTYGPTEETTY-APTEATTYAP 1169

Qy    683 AEPTPKA-LENSPKEPG-----VPTTKTPAATKPEMTTTAKDKTTERDLRTTP----ETT 732
            | || |  | :  ||       ||  :|  |   | |    ::||      ||    | |
Db   1170 TEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPTEETTYAPTEETPYEPTEET 1229

Qy    733 TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTT----------PFKITT---LKT 779
          | ||     |  || : ||    | | : |   |::||          |   ||    :
Db   1230 TYAP-----TEETTYEPTEETTYAPTEETTYAPTEETTYAPTEETMYAPIDETTYGPTEE 1284

Qy    780 TTLAPKVTTTKKTITTTEIMNKPEETAKPKDRAT--NSKATTPKP-QKPTKAPKKPT--- 833
```

```
Db   1285 TTYAPTEATTYAPTEETPYAPTEETTYEPTGETTYAPTEETTYAPTEETTYAPMEETPYE 1344

Qy    834 ---------STKKP-KTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRP--N 881
Db   1345 PAEESTSTVSTEKPCNTEEFTDEPTDEPTDE--PSDEPTDEPTDEPTDLPTDEPSTPCDN 1402

Qy    882 QTPNSKLVEVNPKSEDAG 899
Db   1403 QGINGIGVENKVRYNNAG 1420
```

RESULT 4
T16251
hypothetical protein F35A5.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 20-Sep-1999 #sequence_revision 20-Sep-1999 #text_change 18-Feb-2000
C;Accession: T16251
R;Leimbach, D.
submitted to the EMBL Data Library, January 1996
A;Description: The sequence of C. elegans cosmid F35A5.
A;Reference number: Z18485
A;Accession: T16251
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1274 <LEI>
A;Cross-references: EMBL:U46675; NID:g1166613; PID:g1166621; PIDN:AAB52641.1; GSPDB:GN00
A;Experimental source: strain Bristol N2; clone F35A5
C;Genetics:
A;Gene: CESP:F35A5.1
A;Map position: X
A;Introns: 1272/2

Query Match 14.1%; Score 802; DB 2; Length 1274;
Best Local Similarity 28.0%; Pred. No. 7.2e-28;
Matches 290; Conservative 102; Mismatches 414; Indels 230; Gaps 56;

```
Qy      6 KNRTKK-KP-----TP-----KPPVVDEAGSGLDNGD----FKVTTPDTSTTQHNKVSTS 50
Db    274 KNPTKKWKPPWEDETPVEEVKEPPVPEKKAPVLKKKDPAPAAKARDPSPSKAAPKKVEPS 333

Qy     51 PKITTAKPI-NPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
Db    334 SPVVPPTPVKNPVKKYKPPWEVDDEPAEEVKKPSAPEKKTPVLKRKEPEPSSTTPSSDPS 393

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAE----TTTKGPA----------------LTTP 148
Db    394 PKKAAPAVKPRDSSPKKATPLQADPKAQEVPPTPVKNPVKKYKPPWEVDDEDPVEEVKQP 453

Qy    149 KEPTPTTP----KEPASTTPKEP----TP-TTIKSAPTTPKEPAPTTTKSAPTTPKEPAP 199
Db    454 EAPAKKTPVLKRKEPAAKDTAKPATSKTPETPEKKDPVKPRDSSPKKVAAKPDSAQAPA- 512

Qy    200 TTTKEPA------------------PTTPKEPAPTTTKEPAPT-------TTKSAPTTP 233
Db    513 TPVKNPVKKWRPPWEDDETPADDVSKPTDAKKTPSLAKKDPAPAKESLKPKADTKAPAKP 572

Qy    234 KEP-----APTTPKKPAPTTPKE----PAPTTPKEPTPTTPKEPAPTTKEPAPTTPK-EP 283
Db    573 RDPSPKKVAPTAPEKKTPVLAKKEPAGPADSKTKEPEKSKPRDPSPKKAVPAKPVPKTEV 632

Qy    284 APTAPKKPAP------TTPKEPAPTTPKEPAPTTTKEP----SPTTPKEPAPTTTKSAPT 333
Db    633 APAAVKKPEPISKPKDTAPKKAEPNSPVVP-PTPVKNPVKKWKPPWEDDDAPAKPVSLPE 691

Qy    334 TTKEPAPTTTKSAPTTP-KEPSPTTTKEPAPTTPK--EPAPTTPKKPAP--TTPKEPAPT 388
Db    692 PEKK-TPVLAKKAPTKPDSEAAADPVSGPSSKDPKLAKKAPVKPRDPSPMKAVPIKPAPK 750

Qy    389 TPKEPAPTTTKKPAPTA-PKEPAPTTPKETAPTTPKKLTPTTP-----EKLAPTTPEKPA 442
Db    751 T--EVPPAVVKKPEPVAKSRDPSPKKAK-AEPNSP--VVPPTPVKNPVKKWKPPWEDDDA 805

Qy    443 PTTPEELAPTTPEEP-----TPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTT-PKE 496
Db    806 PAEPVNVPEPEKKTPVLAKKTPVKPRDPSPKKAVPAKPSTKTDAPPVSVKKPEPVSKPKE 865

Qy    497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT-TTKEPTSTTSDKPAPTTP 555
Db    866 PSPKKAEPNSPVVP----PTPVKNPVKKW-KPPWEDDDEPTEEVKKPSE--PEKKTPVLA 918

Qy    556 KGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPT-------TPKKPAP-----KEL 603
Db    919 K-KEPEKPKD-APKVAAKPRDPSPKKAVPE--KEPAKVAAKPRDLSPKKAIPIPANTQEA 974

Qy    604 APTTTKGPTS------TTSDKP-----APTTPKET------APTTPKEPAPTTPKKPAPT 646
Db    975 PPTPVKNPVKKWKPPWEDDDEPAEPVSAPEPEKKTPVLAKKAPAKPRDP---SPKKAAPV 1031

Qy    647 TPETPPPTTSEVSTPTTTKEPTTIHKSP----DESTPELSA-EP---TPKALENSPKEPG 698
Db   1032 AAK-PDPKIPEV-PPTPVKNPVKKWKPPWEDDDEPSEPVSAPEPEKKTPVLAKKAPTKPA 1089

Qy    699 V-------------PTTKTPAAT------KPEMTTTAKDKTTERDLRTTPETT--TAAPK 737
Db   1090 TKPDSEAAADPVSGPTSKDPKLSKKAPVEKPKPTTDPKDDKLKPSPAKKPEKAPEPAAPK 1149

Qy    738 MTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
Db   1150 KWKPVWDDDPDEPEADFTVPAPSKKPDTEDPADPLG---------GPKTKDPK------- 1193

Qy    798 IMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK--------- 848
Db   1194 -LNKKAPAEKPTEK--------PKPKEVSKEPPKPTEPPKP-AAPKKWKPPWEDDPDEPE 1243

Qy    849 ---TTPTPRKMTSTMP 861
Db   1244 ADFTMPAPKKPDTEDP 1259

RESULT 5
S48478
glucan 1,4-alpha-glucosidase (EC 3.2.1.3) - yeast (Saccharomyces cerevisiae)
N;Alternate names: extracellular glucoamylase; mucin-like protein MUC1; protein YIR01
C;Species: Saccharomyces cerevisiae
C;Date: 10-Sep-1999 #sequence_revision 10-Sep-1999 #text_change 12-Nov-1999
C;Accession: S48478; A26877; B26877; S27281; JC6123
R;Rowley, K.
submitted to the EMBL Data Library, October 1994
A;Reference number: S48478
A;Accession: S48478
A;Molecule type: DNA
A;Residues: 1-1367 <ROW>
A;Cross-references: GB:Z47047; EMBL:Z38061; NID:g603997; PID:g763364; GSPDB:GN00009;
R;Yamashita, I.; Nakamura, M.; Fukui, S.
J. Bacteriol. 169, 2142-2149, 1987
A;Title: Gene fusion is a possible mechanism underlying the evolution of STA1.
A;Reference number: A91831; MUID:87194600
A;Accession: A26877
A;Molecule type: DNA
A;Residues: 1-242 <YAM>
A;Cross-references: EMBL:M16164; NID:g172522; PIDN:AAA35014.1; PID:g172525
A;Accession: B26877
A;Molecule type: DNA
A;Residues: 762-1331 <YA2>
A;Cross-references: EMBL:M16165; NID:g172523; PIDN:AAA35015.1; PID:g172526
R;Pardo, J.M.; Ianez, E.; Zalacain, M.; Claros, M.G.; Jimenez, A.
FEBS Lett. 239, 179-184, 1988
A;Title: Similar short elements in the 5' regions of the STA2 and SGA genes from Sacc
A;Reference number: S27281; MUID:89031230
A;Accession: S27281
A;Molecule type: DNA
A;Residues: 1-31 <PAR>
A;Cross-references: EMBL:X13857; NID:g4551; PIDN:CAA32069.1; PID:g4552
R;Lambrechts, M.G.; Bauer, F.F.; Marmur, J.; Pretorius, I.S.
```

Proc. Natl. Acad. Sci. U.S.A. 93, 8419-8424, 1996
A;Title: Mucl, a mucin-like protein that is regulated by Mss10, is critical for pseudohy
A;Reference number: JC6123; MUID:96323237
A;Accession: JC6123
A;Status: nucleic acid sequence not shown
A;Molecule type: DNA
A;Residues: 1-1367 <LAM>
A;Cross-references: GB:U30626; NID:g1304386; PIDN:AAC49609.1; PID:g1304387
C;Genetics:
A;Gene: SGD:MUC1; STA2; MAL5; DEX2; SGD:S0001458
A;Cross-references: MIPS:YIR019c; SGD:S0001458
A;Map position: 9R
C;Superfamily: yeast glucan 1,4-alpha-glucosidase homolog; glucan 1,4-alpha-glucosidase
C;Keywords: glycosidase; hydrolase; polysaccharide degradation; transmembrane protein
F;5-21/Domain: transmembrane #status predicted <TM1>
F;1350-1366/Domain: transmembrane #status predicted <TM2>

Query Match 13.9%; Score 789.5; DB 1; Length 1367;
Best Local Similarity 28.4%; Pred. No. 2.7e-27;
Matches 301; Conservative 104; Mismatches 483; Indels 171; Gaps 47;

```
Qy     35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
Db    236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy     87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
Db    296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy    145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
Db    345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy    195 KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAP 254
Db    405 SSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTES 453

Qy    255 TTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEP 307
Db    454 SSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSS 509

Qy    308 APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP 363
Db    510 APVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTP 568

Qy    364 ----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETA 418
Db    569 VTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSA 624

Qy    419 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKA 472
Db    625 PVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 679

Qy    473 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAP 531
Db    680 SAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 731

Qy    532 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTL 587
Db    732 SSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 790

Qy    588 KEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKK 642
Db    791 SAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESS 850

Qy    643 PAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP 697
Db    851 SAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYP 906

Qy    698 GVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT 757
Db    907 GSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT 966

Qy    758 --TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDR 811
Db    967 VPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA----- 1021

Qy    812 ATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPR 854
Db   1022 ---SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP- 1077

Qy    855 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHV 914
Db   1078 -SVTTVTNFTPTT-ITTTVCSTGT----------NSAGETTSGCSPKTVTTTVPCST 1122

Qy    915 FMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWML 974
Db   1123 GTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTL 1175

Qy    975 SPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1009
Db   1176 APSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207
```

RESULT 6
S49915
extensin-like protein - maize
C;Species: Zea mays (maize)
C;Date: 05-Mar-1995 #sequence_revision 12-May-1995 #text_change 29-Oct-1999
C;Accession: S49915
R;Rubinstein, A.L.; Broadwater, A.H.; Lowrey, K.; Bedinger, P.A.
submitted to the EMBL Data Library, June 1994
A;Description: Pex genes: pollen-specific genes with extensin-like domains.
A;Reference number: S49915
A;Accession: S49915
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-1188 <RUB>
A;Cross-references: EMBL:Z34465; NID:g600117; PIDN:CAA84230.1; PID:g600118

Query Match 13.8%; Score 784; DB 2; Length 1188;
Best Local Similarity 29.9%; Pred. No. 4e-27;
Matches 230; Conservative 59; Mismatches 359; Indels 122; Gaps 33;

```
Qy     13 PTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPSL----PPN 68
Db    460 PTPHSPPAD---------DYVPPTPPVPGKSPPATSPSPQVQPPAASTPPPSLVKLSPPQ 510

Qy     69 SDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKV 128
Db    511 APVG------------------------SPPPPVKTTSPPAPIG----SPSPPPPVSVV 541

Qy    129 -----LAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPA 183
Db    542 SPPPPVKSPPPPAPVGSPPPPEKSPPPPAPVASPPPPVKSP--PPPTLVASPPPPVKSPP 599

Qy    184 PTTTKSAPTTP--KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTP-KEPAPTT 240
Db    600 PPAPVASPPPPVKSPPPPTPVASPPPPAPVASSPPPMKSPPPPTPVSSPPPPEKSPPPPP 659

Qy    241 PKKPAPTTPKEPAPTT------PKE---PTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
Db    660 PAKSTPPPEEYPTPPTSVKSSPPPEKSLPPPTLIPSPPPQEKPTPPSTPSKP-PSSPEKP 718

Qy    292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
Db    719 SP--PKEPVSSPPQTP----KSSPPPAPVSSPPPTPVSSPPALAPVSSPPSVKSSP---- 768

Qy    352 EPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAP 411
Db    769 PPAPLSSPPPAPQVKSSPPPVQVSSP-PPAPKSSPPLAPVSSPPQVEKTSPPPAPLSSPP 827

Qy    412 TTPKETAP----TTP----KKLTPTTPEKLAPTTPEKPAP-----TTPEELAPTTPEEPT 458
```

```
Db    828 LAPKSSPPHVVVSSPPPVVKSSPPPAPVSSPPLTPKPASPPAHVSSPPEVVKPSTP--PA 885

Qy    459 PTTPEEPAPTTPKAAAPNTPKE-PAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTT 517

Db    886 PTTVISP-PSEPKSSPPPTPVSLPPPIVKSSPPPAMVSSP-PMTPKSSPPPVVVSSPPPT 943

Qy    518 LKE---PA-----PTTPKK---PAPKELAPTTTKE---PTSTTSDKPAPTTPKGTAPTTP 563

Db    944 VKSSPPPAPVSSPPATPKSSPPPAPVNLPPPEVKSSPPPTPVSSPPPA---PKSSPPPAP 1000

Qy    564 -KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPT 622

Db   1001 MSSPPPPEVKSPPPPAPVSSPPPPVKSPPPPAPVSSPP---PPVKSPPPPAPVSSPPPPV 1057

Qy    623 -TPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPEL 681

Db   1058 KSPPPPAPISSPPPPVKSPPPPAPVS--SPPPPVKSPPPPAPVSSPPPPIKSPPPPAPVS 1115

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPET 731

Db   1116 SPPPAP------VKPPSLP-PPAPVSSPPPVVTPAPPKKEEQSLPPPAES 1158


RESULT   7
T30826
nascent polypeptide-associated complex alpha chain, muscle splice form gp220 - mouse
N;Alternate names: alpha-NAC protein
C;Species: Mus musculus (house mouse)
C;Date: 22-Oct-1999 #sequence_revision 22-Oct-1999 #text_change 24-Nov-1999
C;Accession: T30826
R;Yotov, W.V.; St-Arnaud, R.
Genes Dev. 10, 1763-1772, 1996
A;Title: Differential splicing-in of a proline-rich exon converts alphaNAC into a muscle
A;Reference number: Z20889; MUID:96312450
A;Accession: T30826
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-2187 <YOT>
A;Cross-references: EMBL:U48363; NID:g1666688; PID:g1666689; PIDN:AAB18732.1
C;Genetics:
A;Gene: Naca
A;Map position: 10
A;Introns: 24/1; 1996/1; 2024/3; 2050/3; 2099/3; 2142/3; 2183/3
A;Note: differential splicing converts alphaNAC into a tissue-specific DNA-binding activ
C;Keywords: alternative splicing; DNA binding; transcription factor


  Query Match             13.6%;  Score 776;  DB 2;  Length 2187;
  Best Local Similarity   27.7%;  Pred. No. 1.6e-26;
  Matches  288;  Conservative  122;  Mismatches  389;  Indels  240;  Gaps   49;

Qy     34 VTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKE---------TSLTVNKETT 84

Db    806 VQSPKVDPIMSDVTPTSPKKTSATAV-PK-------DTSATLSLKSVPAVTSLSPPKAPV 857

Qy     85 VETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPA 144

Db    858 APSNEATIVPTEIPTSLKNALAAATPKETLATSIPKVTSPS----PQKTPKSVSLKGAPA 913

Qy    145 LTT-----------------PKE-------PTPTTPKEPASTT---------PKEPTPTT 171

Db    914 MTSKKATEIAASKDVSPSQFPKEVPLLQHVPPTSPPKSPVSDTLSGALTSPPPKGP-PAT 972

Qy    172 IKSAPTTPKE-PAPTTTKSAPTTPK---------EPAPTTTKEPAPTTPKEPAPTTTKEP 221

Db    973 LAETPTYPKKSPKPAASKKTPATPSPEGVTAVPLEIPPCSKKAPKTAAPKESSATSSSKR 1032

Qy    222 APTT--TKSAPT--------------------TPKEPAPTTPKK-PAPTTPKEPAPTTP 257

Db   1033 APKTAVSKEIPSKGVTAVPLEISLPLKETSKSATPGEKSASSPKRSPKTAGPKE---TPP 1089

Qy    258 KEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTT--PKEPAP----TTPKEPAPTT 311

Db   1090 GGVTAVPPEISLPPKETPQNATPNESLAASSQKRSPKTSVPKETPPGGVTAMPLEIPSAP 1149

Qy    312 TKEPSPTTPK-----EPAPTTTKSAPTTTKEPAPTTT-KSAPTTPK--------EPSPTT 357

Db   1150 QKAPKTAVPKQIPTPEDAVTILAGSPLSPKKASKTAAPKEAPATPSVGVIAVSGEISPSP 1209

Qy    358 TKEPAPTTPKEPAPTTPKKPAPTT--PKE-----------PAPTTPKEPAP-------T 396

Db   1210 KKTSKTAAPKENSATLPPKRSPKTAAPKETPATSSEGVTAVPSEISPSPPTPASKGVPVT 1269

Qy    397 TTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEK-PAPTTPEELAPTTPE 455

Db   1270 LTPKGAPNALAE-SPASPKKVPKTAAPEETSTTP------SPQKIPKVAGPKEASATPPS 1322

Qy    456 EPTPTT--------PEEPAPTTPKAAAPNTPKEPAPTTPKE-PAPTTPKEPAPTTPKETA 506

Db   1323 KKTPKTAVPKETSAPSEGVTAVPLEIPPSPRKAPKTAAPKETPAPS--PEGATTAPVQIP 1380

Qy    507 PTTPKGTAPTTLKEPAPTTP------------------KKPAPKELAPTTTKEPTS-- 544

Db   1381 PSPRKGSKKAGSKE-TPTTPSPEGVTAAPLEIPISSKKTSKMASPKETLVTPSSKKLSQT 1439

Qy    545 ----TTSDKPAPTTPKGTAPTTPKEPAPTTPKE-PAPTTPKGTAPTTLKEPAPTTPKKPA 599

Db   1440 VGPKETSLEGATAVPLEIPPSHKKAPKTVDPKQVPLTPSPK-DAPTTLAE-SPSSPKK-A 1496

Qy    600 PKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETP-----PPT 654

Db   1497 PKTAAPPSER-VTTVPPEKPA-TPQKASGTTASKVPVPAETQEVAVSSRETPVTPAVPPV 1554

Qy    655 TSEVSTPTTTK------EPTTIHKSPDESTPELSAEPTPKALENSPKE-PGVPTTKTPAA 707

Db   1555 KNPSSHKKTSKTIELKEAPATLPPSPTKSPKIPSSKKAPRT--SAPKEFPASPSIK-PVT 1611

Qy    708 TKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQ 767

Db   1612 T--SLAQTAPPSLQKAPSTTIPKENLAAPAV----LPVSSKSPAAPARASASLSPATAAP 1665

Qy    768 DTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTK 827

Db   1666 QTAPKEATTIPSCKKAAATETPIETSTAPSLEGAPKETSE----TSVSKVLMSSP----- 1716

Qy    828 APKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSK 887

Db   1717 -PKKASSSKRASTLP-------------ATTLPSLKEASVLS---------PTATSSGK 1752

Qy    888 LVEVNPKSEDAGGAEGETP 906

Db   1753 DSHISPVS-DACSTGTTTP 1770


RESULT   8
A35175
mucin 1 precursor, repetitive splice form A [validated] - human
N;Alternate names: breast carcinoma-associated DF3 antigen; core protein KP39; episia
ncreatic mucin; polymorphic epithelial mucin (PEM)
N;Contains: mucin 1 precursor, epithelial tumor antigen splice form; mucin 1 precurso
C;Species: Homo sapiens (man)
C;Date: 20-Apr-2000 #sequence_revision 20-Apr-2000 #text_change 02-Jun-2000
C;Accession: A35175; B35175; A35886; A35887; S10572; S40293; A36735; PX0066; S10218;
R;Ligtenberg, M.J.L.; Vos, H.L.; Gennissen, A.M.C.; Hilkens, J.
J. Biol. Chem. 265, 5573-5578, 1990
A;Title: Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene
A;Reference number: A35175; MUID:90202794
A;Accession: A35175
A;Molecule type: mRNA
A;Residues: 1-952,1033-1344 <LIG1>
A;Cross-references: GB:M32738; GB:J05288; NID:g182121; PIDN:AAA35804.1; PID:g182124;
A;Experimental source: splice form A
A;Note: GenBank entries HUMEPISIA1 and HUMEPISIA2 present only the amino-and carboxyl
A;Accession: B35175
A;Molecule type: mRNA
A;Residues: 1-19,29-952,1033-1344 <LIG2>
```

A;Cross-references: GB:M32739; GB:J05288; NID:g182126; PIDN:AAA35806.1; PID:g182129; GB:
A;Experimental source: splice form B
A;Note: GenBank entries HUMEPISIB1 and HUMEPISIB2 present only the amino-and carboxyl-en
R;Gendler, S.J.; Lancaster, C.A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burchel
J. Biol. Chem. 265, 15286-15293, 1990
A;Title: Molecular cloning and expression of human tumor-associated polymorphic epitheli
A;Reference number: A35886; MUID:90368715
A;Accession: A35886
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-992,1033-1344 <GEN>
A;Cross-references: GB:J05581; NID:g188869; PIDN:AAA59876.1; PID:g188870
A;Note: GenBank entry HUMMUCAB includes one copy of the tandemly repeated sequence
R;Lan, M.S.; Batra, S.K.; Qi, W.N.; Metzgar, R.S.; Hollingsworth, M.A.
J. Biol. Chem. 265, 15294-15299, 1990
A;Title: Cloning and sequencing of a human pancreatic tumor mucin cDNA.
A;Reference number: A35887; MUID:90368716
A;Accession: A35887
A;Status: not compared with conceptual translation
A;Molecule type: mRNA
A;Residues: 1-19,29-1109,'S',1111-1339,'A',1341-1344 <LAN>
A;Cross-references: GB:J05582; NID:g189598; PIDN:AAA60019.1; PID:g189599
A;Note: GenBank entry HUMPANMU contains four fewer copies of the tandemly repeated seque
R;Wreschner, D.H.; Hareuveni, M.; Tsarfaty, I.; Smorodinsky, N.; Horev, J.; Zaretsky, J.
Eur. J. Biochem. 189, 463-473, 1990
A;Title: Human epithelial tumor antigen cDNA sequences. Differential splicing may genera
A;Reference number: S10571; MUID:90276413
A;Accession: S10572
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1344 <WRE>
A;Cross-references: EMBL:X52229; NID:g37053
R;Wreschner, D.H.
submitted to the EMBL Data Library, March 1990
A;Reference number: S40293
A;Accession: S40293
A;Molecule type: mRNA
A;Residues: 1-19,29-155,'P',157-175,'P',177-182,'A',184-212,1033-1037,'A',1039-1344 <WR2
A;Cross-references: EMBL:X52229; NID:g37053; PIDN:CAA36478.1; PID:g37054
R;Abe, M.; Siddiqui, J.; Kufe, D.
Biochem. Biophys. Res. Commun. 165, 644-649, 1989
A;Title: Sequence analysis of the 5' region of the human DF3 breast carcinoma-associated
A;Reference number: A36735; MUID:90088473
A;Accession: A36735
A;Molecule type: mRNA
A;Residues: 1-142,'Q',144-162,'Q',164-168 <ABE>
A;Cross-references: EMBL:M31823; NID:g181542; PIDN:AAA35757.1; PID:g181543
R;Masuzawa, Y.; Miyauchi, T.; Hamanoue, M.; Ando, S.; Yoshida, J.; Takao, S.; Shimazu, H
J. Biochem. 112, 609-615, 1992
A;Title: A novel core protein as well as polymorphic epithelial mucin carry peanut agglu
A;Reference number: JX0235; MUID:93123189
A;Accession: PX0066
A;Molecule type: mRNA
A;Residues: 998-1011,'ES',1014-1017;1018-1032,'T',1034-1037;1038-1057 <MAS>
A;Experimental source: gastric carcinoma cell
R;Zrihan-Licht, S.; Baruch, A.; Elroy-Stein, O.; Keydar, I.; Wreschner, D.H.
FEBS Lett. 356, 130-136, 1994
A;Title: Tyrosine phosphorylation of the MUC1 breast cancer membrane proteins cytokine r
A;Reference number: S51026; MUID:95080414
A;Contents: annotation
A;Note: undetermined tyrosine residues in the carboxyl-terminal non-repetitive region ar
C;Comment: This protein is length polymorphic. Individuals may have between 21 and 125 c
partial repeats. The repeat shown is defined by SmaI nuclease sites.
C;Comment: Serine and threonine residues in the tandem repeat domain are extensively gly
C;Comment: For an alternative splice form without a tandem repeat domain, see PIR:S48146
C;Genetics:
A;Gene: GDB:MUC1; PUM
A;Cross-references: GDB:120705; OMIM:158340
A;Map position: 1q21-1q23
A;Introns: 20/1; 62/3; 1165/3; 1184/2; 1230/1; 1270/3; 1320/3
C;Superfamily: polymorphic epithelial mucin
C;Keywords: alternative splicing; duplication; glycoprotein; phosphoprotein; polymorphis
F;1-1344/Product: mucin 1 precursor, splice form A #status predicted <PREA>
F;1-62/Region: mucin 1 amino-terminal non-repetitive

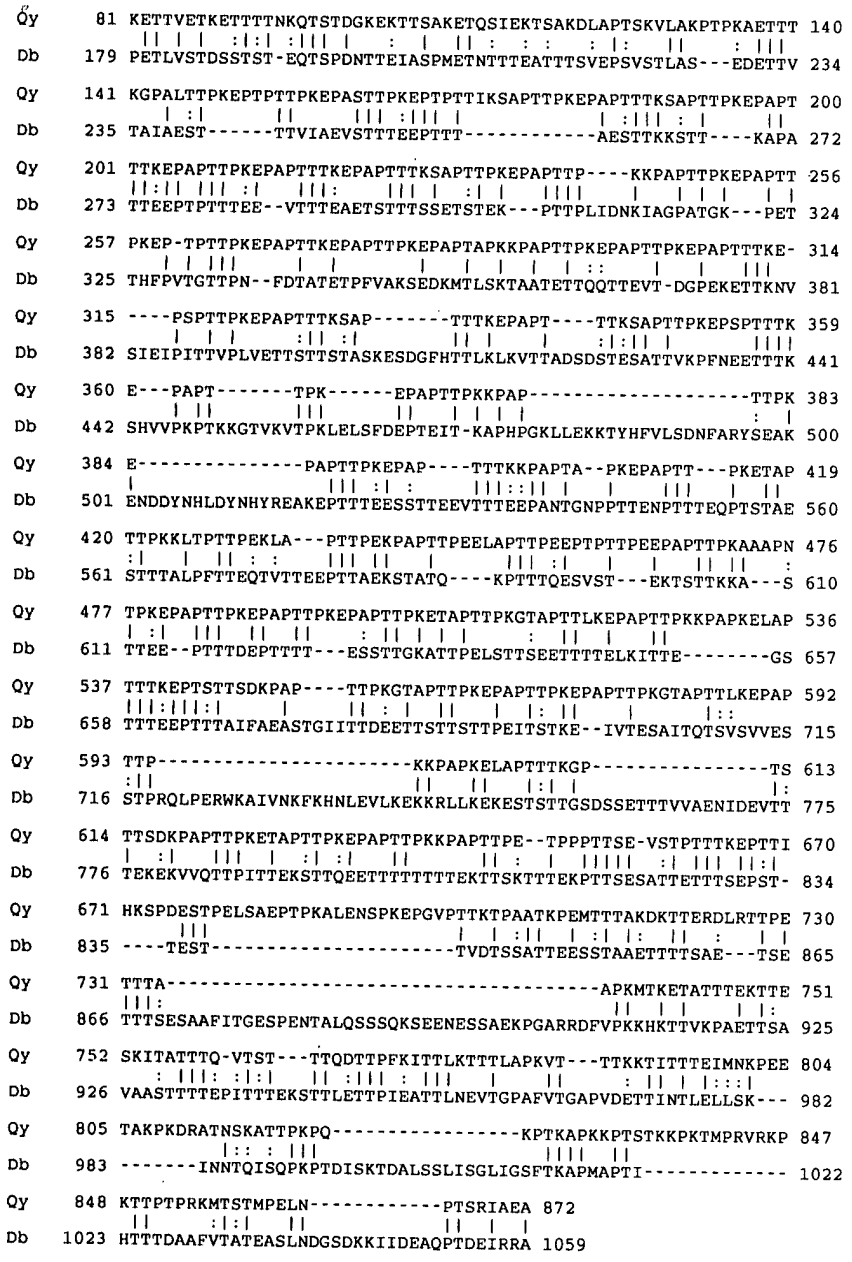
F;1-23/Domain: signal sequence #link PREA #status predicted <SIGA>
F;1-19,29-32/Domain: signal sequence #link PREB #status predicted <SIGB>
F;1-19,29-1344/Product: mucin 1 precursor, splice form B #status predicted <PREB>
F;1-19,29-212,1033-1344/Product: mucin 1 precursor, epithelial tumor antigen splice f
F;138-1017/Region: 20-residue repeats (GSTAPPAHGVTSAPDTRPAP)
F;1143-1344/Region: mucin 1 carboxyl-terminal non-repetitive
F;1245-1272/Domain: transmembrane #status predicted <TRM>
F;1046,1064,1118,1144,1222/Binding site: carbohydrate (Asn) (covalent) #status predic
F;1213/Binding site: phosphate (Tyr) (covalent) #status predicted

Query Match 11.8%; Score 672; DB 1; Length 1344;
Best Local Similarity 27.8%; Pred. No. 3.1e-22;
Matches 310; Conservative 78; Mismatches 472; Indels 256; Gaps 56;

```
Qy     77 LTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
          |||   ||      |  |      :|| | || ||| :  |:  :: |: :: || ||:  :
Db     15 LTVLTATTAPKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 74

Qy    134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTP----------KE 166
          |    :||:|    |    ||              | |     ||||          :
Db     75 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 134

Qy    167 PTPTT-------IKSAPTTPKEPAPTTT-------KSAPTTPKEPAPTTTKEP------A 206
          | | :       : ||| |   ||| :|        ||| |   ||| :|  |      |
Db    135 PAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVTSA 190

Qy    207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP----------KKPAPTTPKEP---- 252
          | |   ||| :|  ||   | ||| |   |  | |           :||| :   |
Db    191 PDT--RPAPGSTAPPAHGVT-SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 247

Qy    253 --APTTPKEPTPTTP----------KEPAPTTKEP-------APTTPKEPAPTAPK---- 289
            || |   |  | |           ||| :  |         || |   |  |||
Db    248 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 307

Qy    290 KPAPTTPKEPAPTTPK----EPAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTT 342
            || |   |  | |         || |   |  | |      |:   |: || :|  ||
Db    308 TSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGV 367

Qy    343 TKSAPTTPKEPSPTTTKEP------APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK- 391
          | ||| |   |:| :|  |       || |   |  | |       || |   |  | | |
Db    368 T-SAPDT--RPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 424

Qy    392 ---EPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPE- 447
             || |   |  |||     |:  :| |       |       || |   |  | |
Db    425 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 484

Qy    448 ---ELAPTTPEEPTPTTP----------EEPAP--TTPKA----AAPNTPKEPAPTTPK- 487
             || |   |  | |             |||  | | |    :||:|   |  | |
Db    485 HGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPA 544

Qy    488 ---EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTP----------KKPAPKEL 534
             || |   |  | |     |     :|| |   |  | |           :|||
Db    545 HGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGST 600

Qy    535 APTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTPKEPAPTTPKG----TAPTTLK 588
          ||      ||    :|||  | |      |:    || |   |  | |       :|| |
Db    601 AP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APDTRPAPGSTAPPAHGVTSAPDTRP 655

Qy    589 EPAPTTP----------KKPAPKELAPTTTKGPTSTTSDKPAP--TTPK----ETAPTTP 632
          | | |           :|||   ||    | ||    :|||  | |       :|| |
Db    656 APGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTR 714

Qy    633 KEPAPTTP----------KKPAPTTPETPPPTTSEVSTPTTTKEPTT-------IHKSPD 675
          | | |             :||| :  | ||     | | |   | :       :  :||
Db    715 PAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 772

Qy    676 ESTPELSAEPTPKALENSPKEPGVPTTKTPAA--------TKPEMTTTAKDK---TTERD 724
            | |    : ::|     | :  | |         |:|   :||      |:  |
Db    773 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPD 832

Qy    725 LRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP--FKIT 775
          |  | :|        |:||    :|       |        |:   |       | |    :|
```

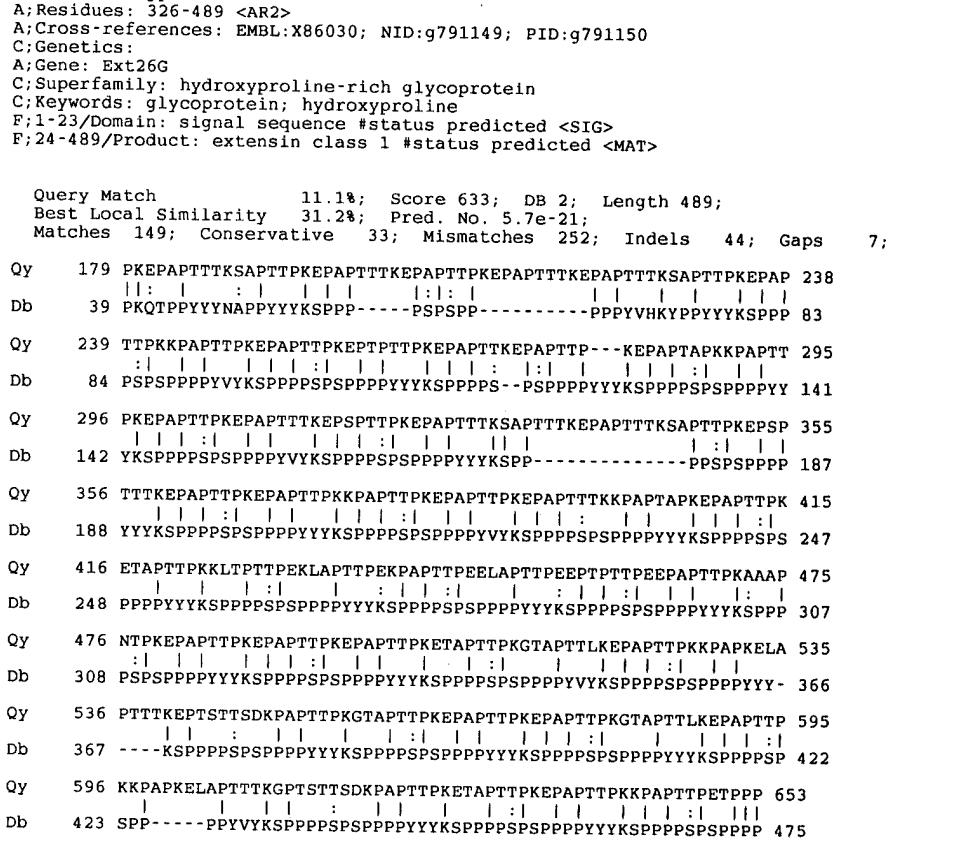
```
Db    833 TRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 888

Qy    776 TLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPT 833
          :   |  ||  |       |:  :    |  || |    |::  | | |   : ||
Db    889 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAP--GSTAPPAHG 946

Qy    834 STKKPKTMPRVRKPKTTPTPRKMTSTMPELNP----TSRIAEAMLQT-TTRP---NQTPN 885
          |  | | |    | :|  |    :: |:  |    |:  |  :     |||   : |
Db    947 VTSAPDTRP---APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP 1003

Qy    886 SKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSDETN 945
          :  |   | :  | |:     |          ||   |  | : :      | : : |:
Db   1004 AHGVTSAPDTRPAPGSTAPPAH---------GVTSAPDNRPALGSTA----PPVHNVTS 1049

Qy    946 ICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 981
            | : |||           :||| ||
Db   1050 ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 1085


RESULT   9
T18535
high molecular mass nuclear antigen - chicken (fragment)
C;Species: Gallus gallus (chicken)
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 15-Oct-1999
C;Accession: T18535
R;Shimada, K.; Harata, M.; Mizuno, S.
J. Cell Sci. 110, 3031-3041, 1997
A;Title: A nuclear matrix-associated high molecular mass nuclear antigen, HMNA, of chick
A;Reference number: Z18955; MUID:9803440
A;Accession: T18535
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-1151 <SHI>
A;Cross-references: EMBL:D88440; NID:d1177138; PID:d1025045; PIDN:BAA24137.1


  Query Match           11.8%;  Score 671;  DB 2;  Length 1151;
  Best Local Similarity  25.9%;  Pred. No. 3e-22;
  Matches  284;  Conservative  100;  Mismatches  419;  Indels  292;  Gaps   55;

Qy     14 TPKP-------PVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTS--PKITTAKPINPRPS 64
          || |       |   :  :|   ||       | :    | :    || |||     ||
Db    113 TPPPSQGPAGTPPPSQGAAGAPKGDGTAQPSGTKSGADGKPAAQDVPKATTA-ATEARP- 170

Qy     65 LPPNSDTSKETSLTVNKETTVETKETTTTNK--QTSTDGKEKTTSAKETQSIEKTSAKDL 122
              :  | || | |   |  |  :    : :||    | :   :||  | :  ::
Db    171 -------ASAASPTVPKATAEATAVTAASQSAPKAATDAAAVTAA---SQSAPKATV-EV 219

Qy    123 APTSKVLAKPTPKAETTTKGPALTTPKEPTP-TTPKEPASTTPKEP----TPTTIKSAPT 177
          | :  :||           |  |   :| | |:|  | |:   :|    :||  |:  |
Db    220 KPAAAAVAKEAKAVTAAAAAPKATAEAKPAPVTSPTIPCSSAEAKPLTAASPTASKA--T 277

Qy    178 TPKEPAPTT----------TKSAPTTPKEPAPTTTK---------------------- 203
             :| | |            | ||:     | | ||
Db    278 AEAKPVPATASLMATKVTAEAKPAPSPSVPKATTDTKAVTATAPKAGPDVKPAVAVCAEA 337

Qy    204 EPAPTTPKEPAPTTTKEPAPTTTKSAPTT------PKEPA-------PTTPKKPAPTTPK 250
          :|||  | :  |      ||| |:  | |       |:  :       |  |:  | | |
Db    338 KPAPPPPQQLPKAAAAAAPTGTELKPATAPPHGSPRANSHTVTVTPPNVPRAAAATVP- 396

Qy    251 EPAPTTPKEPTPTT----PKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKE 306
            |   ||  | ||    |::| |   : || ||  |    |:           || ||::
Db    397 -TAGAVPKASTGTTPAAAPQQPVP---KAAPVTPPSPQQAVPRAATAAA----APVTPQQ 448

Qy    307 P---APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP---TTTKSAPTTPKEPSPTTTKE 360
          |   | |||    |  |   | ||| : | | ::| |   |      |  || ||
Db    449 PVTKAATTTNATPPPQPIPKAATTTTATPVTPQQPIPKAGTDAAPPPAVPKAPSDGRAAT 508

Qy    361 P----APTTPKEPAPTTPKKPAPTTPKEPAP---------TTPKEPAPT-TTKKPAPTAP 406
          |    | | |::| ||    |:  |  :| |          ||  |:|:   | | || |
Db    509 PGVPNAATDPQKPPPTPQSVPSAVTEPKPQPRAAPPPSNEATPAVPSPSPNLKSPLPTIP 568

Qy    407 K-------EPAPTTPK---ETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELA------ 450
```

```
Db    569 KPVPLMALTPQPVTAQMVTQLAATKPSPIVPKASPK-ALMTPPPPPPGLPRALAAAKLLG 627

Qy    451 -PTTP----------EEPTPTTP----EEPAPTTP---------KAAAPNTPKE------ 480
           |::|              | | :|      ||   |          ||:|    |
Db    628 LPSSPVASAMHAKVTPRPLPASPVPMAASPASLGPDAARVALATNAASPGAKPEAAGGNG 687

Qy    481 ----PAPTTPKEPAPTTPKEPAPTTPKETAPT------------TPKGTAPT------- 516
              |  \   : ||      | | |   | |              :| | | |
Db    688 TLMAPMGAANTQMAPIGAAGAAQTAPMGAAHTHVSPMGAGGATQMSPTGAANTHMSPIGA 747

Qy    517 ------------TLKEPAPTTPKKPAPKELAPTTTKEP--------TSTTSDKPAPTTP 555
                      |   |   |  : :|   | ||   |         |:|::      :|
Db    748 GGATQMSPMGAANTQMSPMGATTTQMSPMGAAATTQPSPMGAAATQVTATSAGNTMQVSP 807

Qy    556 KG--TAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEP--APTTPKKPAPKELAPTTTKGP 611
            |  | | ||   | ||| :|:|      | |||  |  | |||:   |  :   ||: |
Db    808 MGAATPPQTPSVGAATTP-QPSPM----GAATTLMSPMGAATTPQ---PSPMGAVTTQPP 859

Qy    612 ----TSTTSDKP-APTTPKET---APTTPKEP---APTTPKKP-APTTPETPP------- 652
              |:||   | | :||: |   | || : |   | ||   |   :||: ||
Db    860 PMAATNTTQPPPMAASTPQSTPMGAATTTQSPPMGATTTQSPPMGASTPQAPPTVAGSPT 919

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTPEL--SAEPTPKALENSPKEPGVPTTKTPAATKP 710
            |      :||   |  : |||      |:   :|  |  |   :   ||| |  :||   |
Db    920 PPPPIPPSPTAQTSPQPMSKSPPPDPPKAPSAAAQTSPAAHVANASPGV-TAVSPA---P 975

Qy    711 EMTTTAKDKTTERDLRTTPETTTAAPKMT-KETATTTEKTTESKITATTTQVTSTTTQDT 769
            | |          |   |   |  || :   ||  ||  |:  :||  | | :    :
Db    976 IGVTEASPSADGARLSPGPTAATDGPKASPAATADVTEAATD--VTAAATAVPA----EA 1029

Qy    770 TPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAP 829
            |          | | : :::   : :::   :    ::   |  ::| ::   |  |
Db   1030 AP---------TKAKRSSSSSSSSSSSSSSSSSSSSSSSDSDSSSSSSESNPASPA--- 1077

Qy    830 KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLV 889
                      | |   :   ||   |  |       : || :|
Db   1078 ------------PAVGDGQQQMTPGAAQSVPP-------VTEAAVQ-------------- 1104

Qy    890 EVNPKSEDAGGAEGE 904
          |    :  | |||  |
Db   1105 EAAAAAAAAAGAERE 1119


RESULT   10
T25697
hypothetical protein F16F9.2 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 15-Oct-1999 #sequence_revision 15-Oct-1999 #text_change 18-Feb-2000
C;Accession: T25697
R;Fulton, B.
submitted to the EMBL Data Library, August 1996
A;Description: The sequence of C. elegans cosmid F16F9.
A;Reference number: Z20071
A;Accession: T25697
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-1229 <FUL>
A;Cross-references: EMBL:U67956; PIDN:AAB07691.1; GSPDB:GN00028; CESP:F16F9.2
A;Experimental source: strain Bristol N2; clone F16F9
C;Genetics:
A;Gene: CESP:F16F9.2
A;Map position: X
A;Introns: 35/1; 361/1; 384/1; 482/2; 517/1; 971/1; 1021/2; 1179/3


  Query Match           11.6%;  Score 659.5;  DB 2;  Length 1229;
  Best Local Similarity  29.0%;  Pred. No. 1e-21;
  Matches  307;  Conservative   78;  Mismatches  339;  Indels  333;  Gaps   52;

Qy     33 KVTT-PDTSTTQ---HNKVSTSPKITTA-KPINPRPSLPPNSDT-------SKETSLTVN 80
          ||||  | |||      | ||:|:| |    || :     :  |       |  | | :
Db    119 KVTTSTDASTTNAPTTGKDSTTPEIITGIVVINSKSESVTDMSTTRFSTTLSPTTELLTS 178
```



RESULT 11
T11622
extensin class 1 precursor - cowpea
C;Species: Vigna unguiculata (cowpea)
C;Date: 16-Jul-1999 #sequence_revision 16-Jul-1999 #text_change 17-Nov-2000
C;Accession: T11622; S54155
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
Mol. Plant Microbe Interact. 10, 95-101, 1997
A;Title: Rhizobia modulate root-hair-specific expression of extensin genes.
A;Reference number: Z17301; MUID:97155574
A;Accession: T11622
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-489 <ARS>
A;Cross-references: EMBL:X91836; NID:g1015936; PIDN:CAA62943.1; PID:g1015937
A;Experimental source: sub_species Red caloona
R;Arsenijevic-Maksimovic, I.; Broughton, W.J.; Krause, A.
submitted to the EMBL Data Library, April 1995
A;Description: A class of root-hair specific extensins involved in rhizobium/legume i
A;Reference number: S54155
A;Accession: S54155
A;Status: preliminary
A;Molecule type: mRNA
A;Residues: 326-489 <AR2>
A;Cross-references: EMBL:X86030; NID:g791149; PID:g791150
C;Genetics:
A;Gene: Ext26G
C;Superfamily: hydroxyproline-rich glycoprotein
C;Keywords: glycoprotein; hydroxyproline
F;1-23/Domain: signal sequence #status predicted <SIG>
F;24-489/Product: extensin class 1 #status predicted <MAT>

Query Match 11.1%; Score 633; DB 2; Length 489;
Best Local Similarity 31.2%; Pred. No. 5.7e-21;
Matches 149; Conservative 33; Mismatches 252; Indels 44; Gaps 7;



RESULT 12
I38346
elastic titin - human (fragment)
C;Species: Homo sapiens (man)
C;Date: 29-May-1998 #sequence_revision 29-May-1998 #text_change 21-Jul-2000
C;Accession: I38346

```
R;Labeit, S.; Kolmerer, B.
Science 270, 293-296, 1995
A;Title: Titins: giant proteins in charge of muscle ultrastructure and elasticity.
A;Reference number: A57430; MUID:96026330
A;Accession: I38346
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-7962 <RES>
A;Cross-references: EMBL:X90569; NID:g1017426; PIDN:CAA62189.1; PID:g1017427
C;Genetics:
A;Gene: GDB:TTN
A;Cross-references: GDB:127867; OMIM:188840
A;Map position: 2q31-2q31

  Query Match           11.1%;  Score 632.5;  DB 2;  Length 7962;
  Best Local Similarity  22.9%;  Pred. No. 9e-20;
  Matches  295;  Conservative  105;  Mismatches  453;  Indels  437;  Gaps   60;

Qy      10 KKKPTP----KPPV---VDEAGSGLDNGDFKVT-----------TPDTSTTQHNKVSTS 50
Db    5837 KKAPPPPKVIKKPVIEKIEKTSRRMEEEKVQVTKVPEVSKKIVPQKPSRTPVQEEVIEVK 5896

Qy      51 -PKITTAKPINPRPSLPPNSDTSKETSLT---VNKETTVETKETTTTNKQTSTDGKEKTT 106
Db    5897 VPAVHTKKMVISEEKMFFASHTEEEVSVTVPEVQKEIVTEEKIHVAVSKRVE-------P 5949

Qy     107 SAKETQSIEKTSAKDLAPT--SKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPAS--- 161
Db    5950 PPKVPELPEKPAPEEVAPVPIPKKVEPPAPKVPEVPKKP-VPEEKKPVPVPKKEPAAPPK 6008

Qy     162 --TTPKEPTPTTIKSAPTTPKEPAP-----------TTTKSAPTTPKEPAPTTTKEPA- 206
Db    6009 VPEVPKKPVPEEKIPVPVAKKKEAPPAKVPEVQKGVVTEEKITIVTQREESP----PPAV 6064

Qy     207 PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTP-KEPAPTTPKEPTPTT- 264
Db    6065 PEIPKKKVPEERK-PVPRKEEEVPPPPKVPA--LPKKPVPEEKVAVPVPVAKKAPPPRAE 6121

Qy     265 ------------------------------------------------------------P 265
Db    6122 VSKKTVVEEKRFVAEEKLSFAVPQRVEVTRHEVSAEEEWSYSEEEEGVSISVYREEEREE 6181

Qy     266 KEPAPTTK-------------------------EPAPTT----------PKEPA------ 284
Db    6182 EEEAEVTEYEVMEEPEEYVVEEKLHIISKRVEAEPAEVTERQEKKIVLKPKIPAKIEEPP 6241

Qy     285 ----PTAPKK-------PAPTTPKE--PAPTTPKEP-APTTTKEPSPTTPKE----PAPT 326
Db    6242 PAKVPEAPKKIVPEKKVPAPVPKKEKVPPPKVPEEPKKPVPEKKVPPKVIKMEEPLPAKV 6301

Qy     327 TTKSAPTTTKEP--APTTTKSAP---TTPKEPSPTTTKE-------------PAPTT--- 365
Db    6302 TEKHMQITQEEKVLVAVTKKEAPPKARVPEEPKRAVPEEKVLKLKPKREEEPPAKVTEFR 6361

Qy     366 -------------PK-EPAP---------------------------------------- 371
Db    6362 KRVVKEEKVSIEAPKREPQPIKEVTIMEEKERAYTLEEEAVSVQREEEYEEYEEYDYKEF 6421

Qy     372 ------------------------------TTPKKPAPT--TPKEPAPTTPKEPAPTTTK 399
Db    6422 EEYEPTEEYDQYEEYEEREYERYEEHEEYITEPEKPIPVKPVPEEPVPTKPKAPPAKVLK 6481

Qy     400 KPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK-------LAPTTPEKPAPTTPEELAPT 452
Db    6482 KAVP-EEKVPVP-IPKKLKPPPPK--VPEEPKKVFEEKIHISITKREKEQVTEP---AAK 6534

Qy     453 TPEEPTPTTPEEPAPTTPKAAAP--NTPKEPAPTTPKEPA------------------- 490
Db    6535 VPMKPKRVVAEEKVPVPRKEVAPPVRVPEVPKELEPEEVAFEEEVVTHVEEYLVEEEEEY 6594

Qy     491 ------PTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTS 544
Db    6595 IHEEEEFITEEEVVPVIPVK-VPEVPRKPVPEE-KKPVPVPKKKEAPPAKVPEVPKKPEE 6652

Qy     545 TT------SDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPK 596
Db    6653 KVPVLIPKKEKP----PPAKVPEVPKKPVP-EEKVPVPVPKKVEAP-----PAKVPEVPK 6702

Qy     597 KPAPKEL----APTTTKGPTSTTSDKPAPTTPKETAPT-TPKE---PAPTTPKKPAPTTP 648
Db    6703 KPVPEKKVPVPAPKKVEAPPAKVPEVPKKLIPEEKKPTPVPKKVEAPPPKVPKKREPVPV 6762

Qy     649 ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAAT 708
Db    6763 PVALPQEEEVLFEEEIVPEEEVLPEEEEVLPE-EEEVLPEEEEVLPEEEEIPPEEEEVPP 6821

Qy     709 KPEMTTTAKDKTTERDL--RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTT 766
Db    6822 EEEYVPEEEEFVPEEEVLPEVKPKVPVPAP-----VPEIKKKVTEKKVVIPKKEEAPPAK 6876

Qy     767 QDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPT 826
Db    6877 VPEVPKKVE--EKRIILPK----EEEVLPVEVTEEPEE----------------EPISEE 6914

Qy     827 KAPKKPTSTKK-----PKTMPRVRK---PKT-TPTPRKMTSTMPELNPTSRIAEAMLQTT 877
Db    6915 EIPEEPPSIEEVEEVAPPRVPEVIKKAVPEAPTPVPKKVEA--PPAKVSKKIPEEKVPVP 6972

Qy     878 TRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIIN 937
Db    6973 VQKKEAPPAKVPEV----------------------------PKKVPEKKVL--VPKKEAV-- 7003

Qy     938 PMLSDETNICNGKPVDGLTTLRNGTLVAFR 967
Db    7004 ------------PPAKGRTVLEEKVSVAFR 7021


RESULT  13
C84672
hypothetical protein At2g27380 [imported] - Arabidopsis thaliana
C;Species: Arabidopsis thaliana (mouse-ear cress)
C;Date: 02-Feb-2001 #sequence_revision 02-Feb-2001 #text_change 02-Feb-2001
C;Accession: C84672
R;Lin, X.; Kaul, S.; Rounsley, S.D.; Shea, T.P.; Benito, M.I.; Town, C.D.; Fujii, C.Y
M.; Koo, H.; Moffat, K.S.; Cronin, L.A.; Shen, M.; VanAken, S.E.; Umayam, L.; Tallon,
euss, D.; Nierman, W.C.; White, O.; Eisen, J.A.; Salzberg, S.L.; Fraser, C.M.; Venter
Nature 402, 761-768, 1999
A;Title: Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana.
A;Reference number: A84420; MUID:20083487
A;Accession: C84672
A;Status: preliminary
A;Molecule type: DNA
A;Residues: 1-761 <STO>
A;Cross-references: GB:AE002093; NID:g5306260; PIDN:AAD41992.1; GSPDB:GN00139
C;Genetics:
A;Gene: At2g27380
A;Map position: 2

  Query Match           11.1%;  Score 632;  DB 2;  Length 761;
  Best Local Similarity  31.1%;  Pred. No. 9.7e-21;
  Matches  217;  Conservative   44;  Mismatches  354;  Indels   82;  Gaps   33;

Qy     132 PTPKAETTTKGPALTTP---KEPTPT-TPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTT 187
Db      69 PPPIQKPPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPT 128

Qy     188 KSAPTTPKEPAPTTTKEPAP--TTPKEPAPTTTKEPAPTTTKSAPTTP----KEPAPT-- 239
Db     129 YSPPIYP----PPIQKPPTPSYSPPVKPPPV---QMPPTPTYSPPIKPPPVHKPPTPTYS 181

Qy     240 TPKKP---APTTPKEPAPTTP----KEPTP--TTPKEPAPTTKEPAPT-TPKEPAPTAPK 289
Db     182 PPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHK 241
```

```
Qy    290 KPAPT-TPKEPAPTTPKEPAP---TTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKS 345
Db    242 PPTPIYSPPIKPPPVHKPPTPIYSPPVKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPP 301

Qy    346 APTTPKEPSPTTTKEP---APTTPKEPAPT--TPKKPAPTTPKEPAPTTPKEPAPTTTKK 400
Db    302 VKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPVKP-PPVHKP 360

Qy    401 PAP--TAPKEPAPT----TPKETAPTTP----KKLTPTTPEKLAPTTPEKPAPTT---PE 447
Db    361 PTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPI 420

Qy    448 ELAPTTPEEPTPTTPEEPAPT----TPKAAAP-NTPKEPAPTTPKEPAPTTPKEPAPTTP 502
Db    421 KLPPVKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTP 480

Qy    503 KETAPTTPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
Db    481 TYSPPVQP----PPVQKPPTPTYSPPVKPPPIQKPPT----PTYSPPIKPPPVKPPTPTY 532

Qy    561 TTPKEPAPTTPKEPAPT-TPKGTAPTTLKEPAPT--TPKKPAPKELAPTTTKGPTSTTSD 617
Db    533 SPPIKP-PPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPP 591

Qy    618 KPAPTTPKETAPTTPKEPAPTTPKKPAPT-TPETPPPTTSEVSTPTTTK--EPTTIHKSP 674
Db    592 VHKPPTPTYSPPIKP----PPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPP 647

Qy    675 DESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTA 734
Db    648 ---TPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPP-TPTYSPPVKPPPVQVPPTPTYS 703

Qy    735 APKMTKETATTTEKTTESKITATTTQVTSTTTQDTTP 771
Db    704 PPVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPP 740
```

RESULT 14
T34513
hypothetical protein ZK783.1 - Caenorhabditis elegans
C;Species: Caenorhabditis elegans
C;Date: 29-Oct-1999 #sequence_revision 29-Oct-1999 #text_change 29-Oct-1999
C;Accession: T34513
R;Favello, A.; Vaudin, M.
submitted to the EMBL Data Library, August 1994
A;Description: The sequence of C. elegans cosmid ZK783.
A;Reference number: Z21536
A;Accession: T34513
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: DNA
A;Residues: 1-3507 <FAV>
A;Cross-references: EMBL:U13646; PIDN:AAC24418.1; GSPDB:GN00021; CESP:ZK783.1
A;Experimental source: strain Bristol N2; clone ZK783
C;Genetics:
A;Gene: CESP:ZK783.1
A;Map position: 3
A;Introns: 14/1; 48/2; 84/1; 196/3; 303/1; 381/1; 586/1; 605/1; 1175/3; 1207/1; 1409/2; 3504/1

Query Match 11.0%; Score 629; DB 2; Length 3507;
Best Local Similarity 24.1%; Pred. No. 5.7e-20;
Matches 252; Conservative 158; Mismatches 416; Indels 218; Gaps 45;

```
Qy     10 KKKPTPKPPVVDEAGSGLDN-GDFKVTT----PDTSTTQHNKVSTSPKITTA-------K 57
Db   1942 KKQPNREKIEIDEENSSSSNSGQEKPTTKGIVSSTSATSSESTTAEPHVTTSISSTTSTK 2001

Qy     58 PINPRPSLPPN---SDTSKETSLTVNKETTV-ETKETTTTNKQTSTDGKEKTTSAKETQS 113
Db   2002 DMTSSKS-PENVTMSSESPEVSTSSSKSTTASETTVSSTPSESSSSEAPLTSSPATTTEV 2060

Qy    114 IEKTSAKDLAP----TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
Db   2061 ITESSVKSTTPKEESSSEITVKLSSKSPEVTESSVKSSPSTPS-TTSQSVTSTVPETSKS 2119

Qy    170 TTIKS-APTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKS 228
Db   2120 TVLSSEAPVTSTSPTEVHT-SSETKPSLSASSTTGDTNSTTPSTSSLASVKSTSAPEGTS 2178

Qy    229 APTTPKEPAPTTPKKPAPTTPK-EPAPTTPKEPTPTTPKEPAPTTKEP-APTTPKEPAPT 286
Db   2179 ASVAPVKLSSLSPDVSQPSTKTFDATESSTVQASETSSGTSVKSTSEPESHVTKLSITSS 2238

Qy    287 APKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTP--KEPAPTTTKSAPTTTKEPAPTTTK 344
Db   2239 NPSSSVPVTSPKSTPTVPE-----STEQPTSTTPSGQSLTPMNSNSEVLTTSEPHVLSSS 2293

Qy    345 SAP------TTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT- 397
Db   2294 LSPDVSQSSTTPNNLSESSTVE---TPKTSSEVSLNSEEPSTTEAPTTLSPDILSTTTN 2349

Qy    398 --------------------TKKPAPTAPK----------EPAPTTPKETAPTTPKKLTP 427
Db   2350 NLSQSSTVSTEDRSEISSENSEKPT-SAPELVTSSVTHVASSSPDVPTES--SEPDDLTG 2406

Qy    428 TTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP-----------APTTPKAAAPN 476
Db   2407 SSTENIPEASSKQTISSTPTPDTTTASEEPTKSTSMSPDLSTTSNVLSESSTTPESSS-- 2464

Qy    477 TPKEPAPTT------------PKEPAPT---------TPKEPAP-----TTPKETAPTTP 510
Db   2465 --KSPVSSSTEGISVVTSTEFSKVPESTISSVLEEDLTKTTPSPILEETTTASETSEPLT 2522

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPA---------PT-------- 553
Db   2523 EDSLTVSVRIHELTTSSENVPKESESTTTSSESSKPSQEPAGILTSTVVVPTSSVSLITA 2582

Qy    554 ---------TP--KGTAP--TTPK-------EPAPTTPKEPAPTTPKGTAPTTLKEPAPT 593
Db   2583 SEIEAITSNTPFKQGRTPITTSPKSLVKSTTSPSTVTSSEPSESTKRTTVSTTVSTTTPT 2642

Qy    594 TPKKPAPKELAPTTTKGPT-STTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPP 652
Db   2643 EETTTSESLILTAAPSKPTESTTESSEAPTTPAKTSETKPSNVSSTSRKSTENVETSTSQ 2702

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTP--ELSAEPTPKALE--NSPKEPGVPTTKTPAAT 708
Db   2703 SGSLESSTMSSTS------SEPETNAPAVTVSSEASSTTLEENSSTSSP----TSSEASV 2752

Qy    709 K-----PEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTT---- 759
Db   2753 KLSSLFPESITSEAVTVSSR----APAEITMSSESHREISTVSSEPSEPEIPLSTTVSPN 2808

Qy    760 ----------QVTSTTTQDTTP-FKITT-----LKTTTLAPKVTTTKKTITTTEIMNKP 802
Db   2809 VVTASSIPSEEPILSSVTSSSTPRVRLITGTPDDLIVSVTVPSHGNRRQNITASSV---P 2865

Qy    803 EETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPK---TTPTPRKMTST 859
Db   2866 SNSTSPIILPSES-LTTPQPPPTTTTTAKPATTSGKRGPPSIQPPAEMFTTPAP------ 2918

Qy    860 MPELN-----PTSRIAEAMLQTTT 878
Db   2919 PPPSNGGYGEETNQEEEQVTSTTT 2942

RESULT 15
I51618
nucleolar phosphoprotein - African clawed frog
C;Species: Xenopus laevis (African clawed frog)
C;Date: 13-Sep-1996 #sequence_revision 13-Sep-1996 #text_change 24-Sep-1999
C;Accession: I51618; S57757
R;Cairns, C.; McStay, B.
J. Cell Sci. 108, 3339-3347, 1995
A;Title: Identification and cDNA cloning of a Xenopus nucleolar phosphoprotein, xNopp
```

A;Reference number: I51618; MUID:96019267
A;Accession: I51618
A;Status: preliminary; translated from GB/EMBL/DDBJ
A;Molecule type: mRNA
A;Residues: 1-990 <CAI>
A;Cross-references: EMBL:X88927; NID:g895920; PIDN:CAA61368.1; PID:g895921
C;Genetics:
A;Gene: xNopp180
C;Superfamily: nucleolus-cytoplasm shuttle phosphoprotein
C;Keywords: phosphoprotein

Query Match 11.0%; Score 626.5; DB 2; Length 990;
Best Local Similarity 26.2%; Pred. No. 2.2e-20;
Matches 253; Conservative 132; Mismatches 393; Indels 187; Gaps 45;

```
Qy      3 DNKKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPR 62
          | || :      ||     |: | |:   :   |     |        |    | : |:
Db     59 DAKKRKPPANGLPKKKSAKESSSSEDSSSEEDEPPAKKRAQ----PAGGKKPVVKAVQPK 114

Qy     63 PSLPPNSDTSKETSLTVNKETTVE--TKETTTTNK----QTSTDGKEKTTSAKETQSIEK 116
           :   : |:| |:    ::| | :   |    | |    :| |  | |::|: |: | |
Db    115 KAKSSSEDSSDESD---SEEETKKPPAKRPAQTPKVAAVKTPTQKKAKSSSS-ESSSSED 170

Qy    117 TSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKEPTPTTI 172
           ::|   }  ||     ||      | |    |    :: ::    ||  |    || |
Db    171 EASKKKQPVIKV----PPKQAVVKAGLASNNGKTADSSSSEDSDSPPAKKTAATKTPPT- 225

Qy    173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTT-TKEPAPTT-----T 226
             | |  :|    |    :: :| :  :: |   |   :| |   :  | ||:     |
Db    226 --KPATAAKPQAKKTAGKKSSSREDSSDSSDEEQKTAKSKPKPDVYSAVPPPTSVSKKKT 283

Qy    227 KSAPTTPKEPAPT----------TPKKPAPTTPKEPAPTTP----KEPTPTTPKEPAPTT 272
           | | |  :|  :           | | |   | : | : |    |:   :|  |   ::
Db    284 LSQPGTKAKPESSDSSDSSDEEEQPAKKAKIVPAKAAASAPKPLAKKAETSTDSESDSSS 343

Qy    273 KEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTP----KEPAP----------------TT 311
           ::   :: |     ||||  ||  |   | :||    |: ||                ||
Db    344 EDEKKSSVKLGVKAAPKK-APAAP--DAKSTPVAAAKKSAPAKKASSSSDSDSSSNEETT 400

Qy    312 TKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA- 370
          ||  : ||| : | | |   ||  |    | |   |  |  :| |: : ::    :
Db    401 TKPAAKTTPAKSAATPTSKTPTNGK---ATPTSKTPAKPGTPKTSTAKKDSSSSDSSDSS 457

Qy    371 ---PTTPKKPAPTTPKEPAPT-TPKEPAPTTTKKPAPTAPKEPA-PTTPKETA------- 418
              || |  | |||  : | | | | |   |  |  ||: | || | ||| :|
Db    458 SDEETTTKPAAKTTPAKSAATPTSKTP---TNSKATPTSKKTPAKPGTPKTSAAKKDSSS 514

Qy    419 -----PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP-KA 472
                :: :| ||   :: | ||| |||       | ||| :|        | ||| |
Db    515 SDSSDSSSDEKKTPA--KRAAKTTPAKPA-------AKTTPAKPA-------AKTTPAKP 558

Qy    473 AAPNTPKEPAPT----TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKEPAPTTPK 527
          || :|| :  ||    :  : : :: ::    :: |    ||| | |:        |
Db    559 AAKSTPGKQVPTKKESSSSDSSDSSSEDEKKSSAKPAVKTTPGKATS-------KPVVAS 611

Qy    528 KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT--PKEPAPTTPKGTAPT 585
          || | :      |  :|: ||     | | | | |   ||  |  ||:   ::   :
Db    612 KPVPAK------KASSSSDSDSSEEETTKTTKPLTKLSPAVKTLPPKKAESSSDSSSDSD 665

Qy    586 TLKEPAPTTPKKPAPKELAPTTTKGP--------TSTTSDKPAPTTPKETAPT--TPKEP 635
          : |:   | | ||  |   |  || |         : : || :    |   ||  :|
Db    666 SEKK---TKPAKPPAKSATPVNTKAPAQNKASKASCSDSDSSSEEEGKSKQPTGKSPAAK 722

Qy    636 APTTPKK-PAPTTPETPPPTTSEVSTPTTTKE-PTTIHKSPDESTPELSAEPTPKALENS 693
          |   ||| |      : |  ::|  |:    |: |     : |      :|:|||   :|
Db    723 ATAPPKKNPVAVNKDKPSSSSSSDSSGDDEKQKPKQAAAAKDVKQGAKAAKPTPKKAASS 782

Qy    694 PKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESK 753
            |    ::     :| : | ||  |:      |||:   || | :   :: :|   :
Db    783 SSE---DSSSDEDVSKAKKTNTAVSKSPV----TTPKAVPAAKKESSSESSDSEDEKQGG 835

Qy    754 ITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTK--KTITTTEIMNKPEETA-KPKD 810
                :||:: ::|                  ||    :  :  :::|   |   |: | |
Db    836 KNTSTTKIANST-----------------PKAAAAECSEESSSSEDEGKANGTSGKRKR 877

Qy    811 RATNS---KATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTP-RKMTSTMPELNPT 866
            :| :   :| ||:          |    | | | |:| | :   || |::     |:||
Db    878 ESTGNAECEAVTPE--------NKKLKAKSPNTFPKVNKKELKNTPFRRVVEEDIEINP- 928

Qy    867 SRIAE 871
            |:|:
Db    929 -RMAD 932
```

Search completed: April 26, 2002, 16:19:22
Job time: 352 sec

GenCore version 4.5

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:19:35 ; Search time 47.4 Seconds
(without alignments)
823.025 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1263
Perfect score: 5698
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....VAALSTAKYKNWPESVYFFK 1064

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	1168	20.5	5179	1	MUC2_HUMAN	Q02817 homo sapien
2	950	16.7	1664	1	SLP1_CLOTM	Q06852 clostridium
3	789.5	13.9	1367	1	AMYH_YEAST	P08640 saccharomyc
4	651	11.4	1255	1	MUC1_HUMAN	P15941 h mucin 1 p
5	556.5	9.8	875	1	FP1_MYTED	Q25460 mytilus edu
6	555.5	9.7	2700	1	ZAN_HUMAN	Q9y493 homo sapien
7	551	9.7	620	1	EXTN_TOBAC	P13983 nicotiana t
8	533	9.4	1087	1	NFH_MOUSE	P19246 mus musculu
9	530.5	9.3	1162	1	TCNA_TRYCR	P23253 trypanosoma
10	530	9.3	865	1	CPN_DROME	Q02910 drosophila
11	518.5	9.1	872	1	FP1_MYTCO	Q25434 mytilus cor
12	509.5	8.9	662	1	MUC1_XENLA	Q05049 xenopus lae
13	503.5	8.8	1970	1	RPB1_HUMAN	P24928 homo sapien
14	498.5	8.7	1970	1	RPB1_MOUSE	P08775 mus musculu
15	497	8.7	831	1	NFH_RAT	P16884 rattus norv
16	493.5	8.7	467	1	RPB1_CRIGR	P11414 cricetulus
17	488.5	8.6	826	1	SSP2_PLAYO	Q01443 plasmodium
18	488	8.6	1020	1	NFH_HUMAN	P12036 homo sapien
19	475.5	8.3	267	1	EXTN_MAIZE	P14918 zea mays (m
20	471.5	8.3	5376	1	ZAN_MOUSE	O88799 mus musculu
21	468.5	8.2	634	1	HWP1_CANAL	P46593 candida alb
22	467	8.2	817	1	VRP1_YEAST	P37370 saccharomyc
23	463	8.1	2142	1	BAT2_HUMAN	P48634 homo sapien
24	454	8.0	797	1	VGLX_HSVEB	P28968 equine herp
25	454	8.0	1161	1	YJ9P_YEAST	P47179 saccharomyc
26	448.5	7.9	670	1	VG50_HSVI1	Q00130 ictalurid h
27	442.5	7.8	751	1	FP1_MYTGA	Q27409 mytilus gal
28	439.5	7.7	1083	1	T2D3_HUMAN	O00268 homo sapien
29	432.5	7.6	439	1	XP2_XENLA	P17437 xenopus lae
30	432	7.6	3164	1	TEGU_HSV11	P10220 herpes simp
31	426	7.5	2715	1	TRX2_HUMAN	Q9umn6 homo sapien
32	424.5	7.4	1125	1	MAP4_MOUSE	P27546 mus musculu
33	424	7.4	307	1	SGS3_DROME	P02840 drosophila
34	424	7.4	1251	1	YQU3_CAEEL	Q09550 caenorhabdi
35	422.5	7.4	1794	1	YAV1_SCHPO	Q10172 schizosacch
36	422.5	7.4	3421	1	TEGU_HSVEB	P28955 equine herp
37	419.5	7.4	2476	1	ZAN_PIG	Q28983 sus scrofa
38	419.5	7.4	2774	1	MAPA_RAT	P34926 rattus norv
39	419	7.4	1185	1	DRPL_HUMAN	P54259 homo sapien
40	415	7.3	1229	1	N121_HUMAN	Q9y2n3 homo sapien
41	411.5	7.2	907	1	VGP3_EBV	P03200 epstein-bar
42	410	7.2	1183	1	DRPL_RAT	P54258 rattus norv
43	407.5	7.2	3256	1	KI67_HUMAN	P46013 homo sapien
44	405.5	7.1	2517	1	NCR2_HUMAN	Q9y618 h nuclear r
45	405	7.1	1411	1	TCOF_HUMAN	Q13428 homo sapien

ALIGNMENTS

```
RESULT   1
MUC2_HUMAN
ID   MUC2_HUMAN     STANDARD;      PRT;  5179 AA.
AC   Q02817; Q14878;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 2 PRECURSOR (INTESTINAL MUCIN 2).
GN   MUC2 OR SMUC.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Intestine;
RX   MEDLINE=94132002; PubMed=8300571;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Siddiki B., Kim Y.S.;
RT   "Molecular cloning of human intestinal mucin (MUC2) cDNA.
RT   Identification of the amino terminus and overall sequence similarity
RT   to prepro-von Willebrand factor.";
RL   J. Biol. Chem. 269:2440-2446(1994).
RN   [2]
RP   SEQUENCE OF 626-1895 AND 4196-5179 FROM N.A.
RC   TISSUE=Colon;
RX   MEDLINE=93016075; PubMed=1400449;
RA   Gum J.R. Jr., Hicks J.W., Toribara N.W., Rothe E.-M., Lagace R.E.,
RA   Kim Y.S.;
RT   "The human MUC2 intestinal mucin has cysteine-rich subdomains located
RT   both upstream and downstream of its central repetitive region.";
RL   J. Biol. Chem. 267:21375-21383(1992).
RN   [3]
RP   SEQUENCE OF 1343-1895 AND 4176-4195 FROM N.A.
RX   MEDLINE=91358717; PubMed=1885763;
RA   Toribara N.W., Gum J.R. Jr., Culhane P.J., Lagace R.E., Hicks J.W.,
RA   Petersen G.M., Kim Y.S.;
RT   "MUC-2 human small intestinal mucin gene structure. Repeated arrays
RT   and polymorphism.";
RL   J. Clin. Invest. 88:1005-1013(1991).
CC   -!- FUNCTION: COATS THE EPITHELIA OF THE INTESTINES, AIRWAYS, AND
CC       OTHER MUCUS MEMBRANE-CONTAINING ORGANS. THOUGHT TO PROVIDE A
CC       PROTECTIVE, LUBRICATING BARRIER AGAINST PARTICLES AND INFECTIOUS
CC       AGENTS AT MUCOSAL SURFACES.
CC   -!- SUBUNIT: MULTIMERIC.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: COLON, SMALL INTESTINE, COLONIC TUMORS,
CC       BRONCHUS, CERVIX AND GALL BLADDER.
CC   -!- PTM: ALL CYSTEINE RESIDUES ARE INVOLVED IN INTRACHAIN OR
CC       INTERCHAIN DISULFIDE BONDS (BY SIMILARITY).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC AND
CC       VARIES AMONG DIFFERENT ALLELES.
CC   -!- SIMILARITY: THE N-TERMINAL DOMAIN SHOWS SOME SIMILARITY TO THAT
CC       OF SILKWORM HEMOCYTIN.
CC   -!- SIMILARITY: CONTAINS 2 VWFC DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 C-TERMINAL CYSTINE KNOT-LIKE DOMAIN (CTCK).
CC   ---------------------------------------------------------------------
```

```
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   -----------------------------------------------------------------------
DR   EMBL; L21998; AAB95295.1; -.
DR   EMBL; M74027; AAA59875.1; -.
DR   EMBL; M94131; AAA59163.1; -.
DR   EMBL; M94132; AAA59164.1; -.
DR   MIM; 158370; -.
DR   InterPro; IPR000359; Cys_knot.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR002400; GF_cysknot.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00007; Cys_knot; 1.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR00438; GFCYSKNOT.
DR   SMART; SM00214; VWC; 2.
DR   SMART; SM00011; VWC_def; 2.
DR   SMART; SM00216; VWD; 4.
DR   PROSITE; PS00022; EGF_1; UNKNOWN_1.
DR   PROSITE; PS01185; CTCK_1; 1.
DR   PROSITE; PS01225; CTCK_2; 1.
DR   PROSITE; PS01208; VWFC; 2.
KW   Glycoprotein; Repeat; Signal.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21   5179       MUCIN 2.
FT   DOMAIN     1401   1747       APPROXIMATE REPEATS.
FT   REPEAT     1401   1416       1.
FT   REPEAT     1417   1432       2.
FT   REPEAT     1433   1448       3.
FT   REPEAT     1449   1464       4.
FT   REPEAT     1465   1471       5.
FT   REPEAT     1472   1478       6.
FT   REPEAT     1479   1494       7A.
FT   REPEAT     1495   1517       7B.
FT   REPEAT     1518   1533       8A.
FT   REPEAT     1534   1556       8B.
FT   REPEAT     1557   1572       9A.
FT   REPEAT     1573   1596       9B.
FT   REPEAT     1597   1612       10A.
FT   REPEAT     1613   1635       10B.
FT   REPEAT     1636   1651       11A.
FT   REPEAT     1652   1675       11B.
FT   REPEAT     1676   1683       12.
FT   REPEAT     1684   1699       13.
FT   REPEAT     1700   1715       14.
FT   REPEAT     1716   1731       15.
FT   REPEAT     1732   1747       16.
FT   DOMAIN     4815   4886       VWFC 1.
FT   DOMAIN     4924   4991       VWFC 2.
FT   DOMAIN     5075   5160       CTCK.
FT   DISULFID   5075   5122       BY SIMILARITY.
FT   DISULFID   5089   5136       BY SIMILARITY.
FT   DISULFID   5098   5152       BY SIMILARITY.
FT   DISULFID   5102   5154       BY SIMILARITY.
FT   DISULFID      ?   5159       BY SIMILARITY.
FT   CARBOHYD    163    163       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    423    423       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    670    670       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    894    894       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1139   1139       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1154   1154       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1215   1215       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1230   1230       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1246   1246       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1787   1787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1820   1820       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4339   4339       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4351   4351       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4362   4362       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4373   4373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4422   4422       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4438   4438       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4502   4502       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4616   4616       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4627   4627       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4752   4752       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4787   4787       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4881   4881       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4888   4888       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4955   4955       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   4970   4970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5019   5019       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5038   5038       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   5069   5069       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   1351   1351       H -> L (IN REF. 3).
FT   CONFLICT   1412   1412       T -> S (IN REF. 3).
FT   CONFLICT   1449   1449       L -> P (IN REF. 3).
FT   CONFLICT   1504   1504       M -> T (IN REF. 3).
FT   CONFLICT   4192   4192       G -> S (IN REF. 2).
SQ   SEQUENCE   5179 AA;  540295 MW;  85CD7571FB9A5663 CRC64;
```

```
  Query Match           20.5%;  Score 1168;  DB 1;  Length 5179;
  Best Local Similarity 32.5%;  Pred. No. 1.9e-40;
  Matches  363;  Conservative   53;  Mismatches  448;  Indels  254;  Gaps   40;

Qy      21 DEAGSGLDNGDFK----------------VTTPDTSTTQH-NKVSTSPK---------- 52
           |   || |:|| :                |  |  |  ||  ||
Db    1312 DHPSSGSDDGDREPFDGVCGAPEDIECRSVKDPHLSLEQHGQKVQCDVSVGFICKNEDQF 1371

Qy      53 ------------------------ITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETK 88
                                    |||  |    || || : |:   : | :  ||
Db    1372 GNGPFGLCYDYKIRVNCCWPMDKCITTPSPPTTTPSPPPTTTTTLPPTTTPSPPTT---- 1427

Qy      89 ETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTP 148
            ||||   |:|     ||:                ||:     |:|   |||  |  |||
Db    1428 -TTTTPPPTTTPSPPITTTTTP-----------LPTT----TPSPPISTTTTPPPTTTP 1470

Qy     149 KEPT-----PTTPKEPASTTPKEPTPTTIKSAP-TTPKEP-------APTTTKSAP-TTP 194
             ||    |||   | :||   | |||  | | |||  |         |||| | | ||
Db    1471 SPPTTTPSPPTTTPSPPTTTTTTPPPTTTPSPPMTTPITPPASTTTLPPTTTPSPPTTTT 1530

Qy     195 KEPAPTTTKEPAPTTPKEP-------APTTTKEPAPTTTKSAP--TTPKEPAPTTPKKPA 245
             | ||||  |  |||  |         ||||  | |||| : |  |||  |  |||  |
Db    1531 TTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTTPPPTTTPSPPTTTTPSPPT 1590

Qy     246 PTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEP-------APTAPKKPAPTTPKE 298
            ||   |  |||  || ||   |  ||  |  |||  |         ||    | |||
Db    1591 ITTTTPPPTTTPSPPTTTTTTPPPTTTPSPPTTTPITPPTSTTTLPPTTTPSPPPTTTTT 1650

Qy     299 PAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP---TTPKEPSP 355
            | |||   |  |||  |  ||   | |||| |:| ||    |||| : |   |||  |
Db    1651 PPPTTTPSPPTTTTPSPPITTTTTPPPTTTPSSPITTTPSPPTTTMTTPSPTTTPSSPI- 1709

Qy     356 TTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTP- 414
            |||  |: ||   | |||   |:|||   |  ||     ||||  |  | |  |: | |
Db    1710 TTTTTPSSTTTPSPPPTTMTTPSPTTTPSPPTTTMTTLPPTTTSSPLTTTPLPPSITPPT 1769

Qy     415 ----KETAPTTP-------------------------------------------KKLTP 427
                | ||||                                            : |
Db    1770 FSPFSTTTPTTPCVPLCNWTGWLDSGKPNFHKPGGDTELIGDVCGPGWAANISCRATMYP 1829

Qy     428 ------------------------------------------------TTPEKLAPTT 437
                                                           | |  :  ||
Db    1830 DVPIGQLGQTVVCDVSVGLICKNEDQKPGGVIPMAFCLNYEINVQCCECVTQPTTMTTTT 1889

Qy     438 PEKPAPTTPEELAPTTPEEPTPT-TPEEPAPTTPKAAAPNTPKEPAPTTPKEP-----AP 491
            | | | |   :  ||   |||| |  :   |||          | ||  : |
Db    1890 TENPTPPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITT 1949
```

```
Qy     492 TTPKEPAPT-TPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKP 550
           ||   | || |  :|  |||  |  |    | ||  : |    :  |||  || | :
Db    1950 TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQ 2009

Qy     551 AP-TTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 609
           || |||  |  |    | ||  : |  |||  |  |    | ||  : |    :  |||
Db    2010 TPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPTTTPITTTTTV 2068

Qy     610 GPTSTTSDKPAP-TTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPT 668
           ·|| | :    | |||  |  |    | ||  : |  |||  |   | :   ||| |: ||
Db    2069 TPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPT-TTPITTTTTVTPTPTPTGTQTPT 2127

Qy     669 TIHKSPDESTPELSAEPTPKALENSPKEP----------GVPT-TKTPAATKPEMTTTAK 717
           |   :|  :|  ::  |||   :     |            || |:||  |    |||
Db    2128 T---TPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2184

Qy     718 DKTTERDLR---TTP--ETTTAAPKMT---KETATTTEKTTESKITATTTQVTSTTTQDT 769
           |    :  |||  |||  |  |    :| |||  || : :| | |  | | | |  |
Db    2185 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTP-TGTQTPTT 2243

Qy     770 TPFKITTLKTTTLAPKVTTT-KKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKA 828
           ||  |||  |||: |  | |  :| ||| |      |  |    | :  |||     |
Db    2244 TP--ITT--TTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVT 2299

Qy     829 P-KKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNP---TSRIAEAMLQTTTR-PNQT 883
           |   || |: | | |       ||||    :  |   |  |: :      | |: |  |
Db    2300 PTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTTPITTTTTVTPTPTPTGTQTPTTT 2359

Qy     884 PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTP 921
           | :    | |    |    :||     |      |||
Db    2360 PITTTTTVTPTPTPTG---TQTP--TTTPITTTTTVTP 2392


RESULT   2
SLP1_CLOTM
ID   SLP1_CLOTM     STANDARD;      PRT;  1664 AA.
AC   Q06852;
DT   01-JUN-1994 (Rel. 29, Created)
DT   01-JUN-1994 (Rel. 29, Last sequence update)
DT   01-FEB-1996 (Rel. 33, Last annotation update)
DE   CELL SURFACE GLYCOPROTEIN 1 PRECURSOR (OUTER LAYER PROTEIN B) (S-LAYER
DE   PROTEIN 1).
GN   OLPB.
OS   Clostridium thermocellum.
OC   Bacteria; Firmicutes; Bacillus/Clostridium group; Clostridiaceae;
OC   Clostridium.
OX   NCBI_TaxID=1515;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=NCIB 10682;
RX   MEDLINE=93209931; PubMed=8458832;
RA   Fujino T., Beguin P., Aubert J.-P.;
RT   "Organization of a Clostridium thermocellum gene cluster encoding the
RT   cellulosomal scaffolding protein CipA and a protein possibly involved
RT   in attachment of the cellulosome to the cell surface.";
RL   J. Bacteriol. 175:1891-1899(1993).
CC   -!- SUBUNIT: ASSEMBLED INTO MONO-LAYERED CRYSTALLINE ARRAYS.
CC   -!- SUBCELLULAR LOCATION: CELL WALL.
CC   -!- SIMILARITY: CONTAINS 4 S-LAYER HOMOLOGY (SLH) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X67506; CAA47841.1; -.
DR   InterPro; IPR001119; SLH.
DR   Pfam; PF00395; SLH; 3.
DR   PROSITE; PS01072; SLH_DOMAIN; 2.
KW   Cell wall; S-layer; Signal; Repeat.
FT   SIGNAL        1     28       POTENTIAL.
FT   CHAIN        29   1664       CELL SURFACE GLYCOPROTEIN 1.
FT   DOMAIN       36    763       4 X 156 AA APPROXIMATE REPEATS.
FT   REPEAT       36    191       1.
FT   REPEAT      207    363       2.
FT   REPEAT      409    565       3.
FT   REPEAT      607    763       4.
FT   DOMAIN      771   1377       APPROXIMATE TANDEM REPEATS OF
FT                                T-P-S-D-E-P.
FT   DOMAIN     1378   1449       GLY/PRO/SER/THR-RICH.
FT   DOMAIN     1453   1494       SLH 1 (INCOMPLETE).
FT   DOMAIN     1495   1565       SLH 2.
FT   DOMAIN     1566   1625       SLH 3.
FT   DOMAIN     1626   1646       SLH 4 (INCOMPLETE).
SQ   SEQUENCE   1664 AA;  178194 MW;  5F396695BA9FE74B CRC64;


  Query Match           16.7%;  Score 950;  DB 1;  Length 1664;
  Best Local Similarity 31.0%;  Pred. No. 4.7e-32;
  Matches  312;  Conservative  104;  Mismatches  361;  Indels  230;  Gaps   55;

Qy     128 VLAKPTP-KAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTT 186
           |: :| | || :    |  |   ||||:           ||||:     ||   || |:
Db     758 VVIQPAPIKAASDEPIPTDTPSDEPTPS----------DEPTPS---DEPTPSDEPTPSD 804

Qy     187 TKSAPTTPKEPAPTTTKEPAPTTPKEPA----PTTTKEPAPTTTKSAPTTPKEPAPTTPK 242
             :   ||:|| || |    ||   ||      || : || |:   :   ||:|| ||
Db     805 EPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTP 864

Qy     243 KPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPT-TPKEPAPTAPKKPAPTTPKEPAP 301
             ||   || |:   ||||:   || | : || |: ||:|| ||       ||   || |
Db     865 SDEPTPSDEPTPS--DEPTPS--DEPTP-SDEPTPSETPEEPIPTDTPSDEPTPSDEPTP 919

Qy     302 TTPKEPA--PTTTKEPSPT-TPKEPAPTTTKS-APTTTKEPAPT---TTKSAPTTPKEPS 354
           :    |:  || : ||:|: ||:|| || | |  || : || |:   |    ||   ||:
Db     920 SDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPT 979

Qy     355 ----PTTTKEPAPT-TPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEP 409
               || : || |: ||:|| ||      ||   || |:    |:   |    ||   ||
Db     980 PSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSDEP 1039

Qy     410 APT---TPKETAPTTPKKLTPTTPEKLAPTTPEKPAPT---TPEELAPTTPEEPTPT-TP 462
           |:   || |    ||:: ||      ||  ::| |:   || :  ||  :||||: ||
Db    1040 TPSDEPTPSE----TPEEPIPTDTPSDEPTPSDEPTPSDEPTPSD-EPTPSDEPTPSETP 1094

Qy     463 EEPAPTTPKAAAPNTPKEPAPTTPKEPA--PTTPKEPAPT-TPKETAPTTPKGTAPTTLK 519
           ||| ||   :  |    || |:    |:  ||   || |: ||:|  ||       ||
Db    1095 EEPIPTDTPSDEPTPSDEPTPSDEPTPSDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD 1154

Qy     520 EPAPTTPKKPAPKELAPTTTKEPT-STTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTT 578
           || |:   :| | :  || .: ||| | | ::| ||      ||   || |:   || |:
Db    1155 EPTPS--DEPTPSD-EPTPSDEPTPSETPEEPIPTDTPSDEPTPSDEPTPS--DEPTPS- 1208

Qy     579 PKGTAPTTLKEPAPT-TPKKPAPKELAPTTTKGPTSTTSDKPAPT---TPKETAPTTPKE 634
             ||   || |: ||::|                || | ||:| |:   || :  ||   |
Db    1209 ---DEPTPSDEPTPSETPEEPI----------PTDTPSDEPTPSDEPTPSD-EPTPSDE 1253

Qy     635 PAPTTPKKPAPT-TPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENS 693
           | |:   :| |: ||| | || :    || : |||      || ||  | ||||
Db    1254 PTPS--DEPTPSETPEEPIPTDTPSDEPTPSDEPT----PSDEPTP--SDEPTPSDEPTP 1305

Qy     694 PKEPGVPTTKTPAAT--KPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTE 751
             ||      ||: |  :|  | |  |: |  |      | :  |  :|  |  :| |
Db    1306 SDEPTPSDEPTPSETPEEPIPTDTPSDEPTPSD----EPTPSDEPTPSDEPTPSDEPTPS 1361

Qy     752 SKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDR 811
           : | :    | | :: ||       |||  |  :||                       |
Db    1362 DEPTPSDEPTPSETPEEPTP-------TTTPTPTPSTT-----------------PTSG 1396

Qy     812 ATNSKAT--------TPKPQKPTKAP-KKPTSTKKPKTMPRVRKPKTTPTPRKM------ 856
           : | :             |  ||  | |||||  |     : :|  :  |  :
Db    1397 SGGSGGSGGGGGGGGGTVPTSPTPTPTSKPTSTPAP---TEIEEPTPSDVPGAIGGEHRA 1453
```

```
Qy     857 -------TSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHM- 908
                  |  || | |    |  :             :||:     :::: ||:  :|:
Db    1454 YLRGYPDGSFRPERNITRAEAAVIF-----------AKLL----GADESYGAQSASPYSD 1498

Qy     909 LLRPH--------------------VFMPE--------VTPDMDYLPRVPNQGI----- 934
           |   |                    | |:          |  : :| :|  | |
Db    1499 LADTHWAAWAIKFATSQGLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLA 1558

Qy     935 ---IINPMLSDETNICNG----KPVDGLTTL------RNGTLVAFRGHYFWMLSPFSPPS 981
             | ||    |    | |    : :: ||:|       :||                | | :
Db    1559 TIDISNPKFDD----CVGHWAQEFIEKLTSLGYISGYPDGT-------------FKPQN 1600

Qy     982 PARRITEVWGIPSPIDTVFTRCNCEGKTFFFKD--SQYWRFTNDIKD 1026
             :|   |    : |:    |    |    | |    || |  || |
Db    1601 YIKRSESV----ALINRALERGPLNGAPKLFPDVNESYWAF-GDIMD 1642
```

```
RESULT   3
AMYH_YEAST
ID   AMYH_YEAST     STANDARD;      PRT;  1367 AA.
AC   P08640; P08068;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-FEB-1995 (Rel. 31, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   GLUCOAMYLASE S1/S2 PRECURSOR (EC 3.2.1.3) (GLUCAN 1,4-ALPHA-
DE   GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE).
GN   STA1 OR STA2 OR MAL5 OR YIR019C.
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=S288C / AB972;
RA   Barrell B.G., Badcock K., Bankier A.T., Bowman S., Brown D.,
RA   Churcher C.M., Connor R., Copsey T., Dear S., Devlin K., Fraser A.,
RA   Gentles S., Hamlyn N., Horsnell T.S., Hunt S., Jagels K., Jones M.,
RA   Louis E., Lye G., Moule S., Moule T., Odell C., Pearson D.,
RA   Rajandream M.A., Riles L., Rowley N., Skelton J., Smith V.,
RA   Walsh S.V., Whitehead S.;
RL   Submitted (DEC-1994) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   SEQUENCE OF 1-242 AND 762-1331 FROM N.A.
RX   MEDLINE=87194600; PubMed=3106330;
RA   Yamashita I., Nakamura M., Fukui S.;
RT   "Gene fusion is a possible mechanism underlying the evolution of
RT   STA1.";
RL   J. Bacteriol. 169:2142-2149(1987).
RN   [3]
RP   SEQUENCE OF 1-31 FROM N.A.
RC   STRAIN=SPX101-1C;
RX   MEDLINE=89031230; PubMed=3141213;
RA   Pardo J.M., Ianez E., Zalacain M., Claros M.G., Jimenez A.;
RT   "Similar short elements in the 5' regions of the STA2 and SGA genes
RT   from Saccharomyces cerevisiae.";
RL   FEBS Lett. 239:179-184(1988).
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF TERMINAL 1,4-LINKED ALPHA-D-
CC       GLUCOSE RESIDUES SUCCESSIVELY FROM NON-REDUCING ENDS OF THE CHAINS
CC       WITH RELEASE OF BETA-D-GLUCOSE.
CC   -!- SIMILARITY: TO S.POMBE SPBC215.13.
CC   -!- SIMILARITY: SOME, TO S.POMBE SPCC285.13C.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; Z38061; CAA86176.1; -.
DR   EMBL; M16164; AAA35014.1; -.
DR   EMBL; M16165; AAA35015.1; -.
DR   EMBL; X13857; CAA32069.1; -.
DR   PIR; B26877; B26877.
DR   PIR; A26877; A26877.
DR   PIR; S48478; S48478.
DR   SGD; S0001458; MUC1.
KW   Hydrolase; Glycosidase; Polysaccharide degradation; Glycoprotein;
KW   Signal; Multigene family.
FT   SIGNAL       1     21       POTENTIAL.
FT   CHAIN       22   1367       GLUCOAMYLASE S1/S2.
FT   DOMAIN     210   1367       SER/THR-RICH.
FT   CARBOHYD   817    817       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   874    874       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1367 AA;  136110 MW;  91C00E2DBD61AA9D CRC64;
```

```
  Query Match           13.9%;  Score 789.5;  DB 1;  Length 1367;
  Best Local Similarity  28.4%;  Pred. No. 1.2e-25;
  Matches  301;  Conservative  104;  Mismatches  483;  Indels  171;  Gaps   47;

Qy      35 TTPDTSTTQHNKVSTSPKITTAKPINP--------RPSLPPNSDTSKETSLTVNKETTVE 86
           :| ::|||  :   :|  :|  | |          :|: |  :  :||     : :||
Db     236 STSESSTTTSSTSESSTSSSTTAPATPTTTSCTKEKPTPPTTTSCTKEKPTPPHHDTTPC 295

Qy      87 TKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETT--TKGPA 144
           ||: |||:| | |   :|||:   | |   |:    ||      ||| : ||  :  |
Db     296 TKKKTTTSK-TCT---KKTTTPVPTPS-SSTTESSSAPV------PTPSSSTTESSSAPV 344

Qy     145 LTTPKE------PTPTTPKEPASTTPKEPTPTTIKSAP----TTPKEPAPTTTKSAPTTP 194
             ::  |      |||::    :|: |   : |   |||    ||    ||  | |: ||
Db     345 TSSTTESSSAPVPTPSSSTTESSSAPVTSSTTESSSAPVTSSTTESSSAPVPTPSSSTTE 404

Qy     195 KEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAP 254
             || |:     ||    || |:   : | : ||| |    : ||    || |
Db     405 SSSAPVTSS----TTESSSAPVTS---STTESSSAPVT----SSTTESSSAPVTSSTTES 453

Qy     255 TTPKEPTP---TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAP----TTPKEP 307
           ::   |||   ||    || |   : ||    ||  |   : ||    ||    ||
Db     454 SSAPVPTPSSSTTESSSAPVT---SSTTESSSAP-VPTPSSSTTESSSAPVTSSTTESSS 509

Qy     308 APTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAP----TTTKSAPTTPKEPSPTTTKEPAP 363
           ||  |   | ||    ||  | |: ||    ||    ||  |:   |  | ||    |
Db     510 APVPTPS-SSTTESSSAPAPTPSSSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSTP 568

Qy     364 ----TTPKEPAPT-TPKKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETA 418
               ||    ||  ||      :   | ||    |: :||:  :  ||   : ||   :|
Db     569 VTSSTTESSSAPVPTPSSSTTESSSAPVPT----PSSSTTESSSAPAPTPSSSTTESSSA 624

Qy     419 PTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPT-TPEEPTPTTPEEPAP-----TTPKA 472
           | |    : ||    ||  |   : ||    ||  ||   |  :   | |     ||  :
Db     625 PVT----SSTTESSSAP-VPTPSSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 679

Qy     473 AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPT-TPKGTAPTTLKEPAPTTPKKPAP 531
           :|| |    : ||    || |    : ||   :||  ||  :   :    | ||
Db     680 SAPVT----SSTTESSSAPVT----SSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTE 731

Qy     532 KELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAP---TTPKEPAPT-TPKGTAPTTL 587
             ||  |   :|||    || |   | ::   | |   ||    ||  ||  :   :
Db     732 SSSAPVPTPS-SSTTESSSAPVTSSTTESSSAPVPTPSSSTTESSSAPVPTPSSSTTESS 790

Qy     588 KEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAP-----TTPKK 642
           | ||          :||  |   :|  :     :||  ::  :   | |     ||
Db     791 SAPVPTPSSSTTESSVAPVPTPSSSSNITSSAPSSTPFSSSTESSSVPVPTPSSSTTESS 850

Qy     643 PAPTTPETPPPTTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPTPKALENSPKEP 697
             || :  |   : : | ||:     |:  |::|   |  || | :  |     :| | |
Db     851 SAPVSSSTTESSVAPVPTPSSSSNITSSAPSSI---PFSSTTESFSTGT-TVTPSSSKYP 906

Qy     698 GVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITAT 757
           |  |  : ::|         |  |:     ||  |||      |     ||   :  :| |
Db     907 GSQTETSVSSTTETTIVPTKTTTSVTTPSTTTITTTVCSTGTNSAGETTSGCSPKTVTTT 966

Qy     758 --TTQVTSTTTQDTTPFKITTLKTTTLAPKVTT---TKKTITTT-EIMNKPEETAKPKDR 811
```

```
Db    967 VPTTTTTSVTTSSTTTITTTVCSTGTNSAGETTSGCSPKTITTTVPCSTSPSETA----- 1021

Qy    812 ATNSKATTPKPQKP----------------TKAPKKPTSTKKPKTMPRVRKPKTTPTPR 854

Db   1022 ---SESTTTSPTTPVTTVVSTTVVTTEYSTSTKPGGEITTTFVTKNIPTTYLTTIAPTP- 1077

Qy    855 KMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHV 914

Db   1078 -SVTTVTNFTPTT-ITTTVCSTGT------------NSAGETTSGCSPKTVTTTVPCST 1122

Qy    915 FMPEVTPDMDYLPRVPNQGIIINPMLSDETNICNGKPVDGLTTLRNGTLVAFRGHYFWML 974

Db   1123 GTGEYTTEATTLVTTAVTTTVVTTESSTGTNSA-GKTTTGYTTKSVPTT------YVTTL 1175

Qy    975 SPFSPPSPARRITEVWGIPSPIDTVFTRC----NCEGKT 1009

Db   1176 APSAPVTPATN-----AVPTTITT--TECSAATNAAGET 1207


RESULT   4
MUC1_HUMAN
ID   MUC1_HUMAN     STANDARD;      PRT;  1255 AA.
AC   P15941; P15942; P13931; P17626; Q14128; Q16442; Q16437; Q9Y4J2;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   MUCIN 1 PRECURSOR (POLYMORPHIC EPITHELIAL MUCIN) (PEM) (PEMT)
DE   (EPISIALIN) (TUMOR-ASSOCIATED MUCIN) (CARCINOMA-ASSOCIATED MUCIN)
DE   (TUMOR-ASSOCIATED EPITHELIAL MEMBRANE ANTIGEN) (EMA) (H23AG) (PEANUT-
DE   REACTIVE URINARY MUCIN) (PUM) (BREAST CARCINOMA-ASSOCIATED ANTIGEN
DE   DF3).
GN   MUC1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Pancreas;
RX   MEDLINE=90368716; PubMed=2394722;
RA   Lan M.S., Batra S.K., Qi W.-N., Metzgar R.S., Hollingsworth M.A.;
RT   "Cloning and sequencing of a human pancreatic tumor mucin cDNA.";
RL   J. Biol. Chem. 265:15294-15299(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=90202794; PubMed=2318825;
RA   Ligtenberg M.J.L., Vos H.L., Gennissen A.M.C., Hilkens J.;
RT   "Episialin, a carcinoma-associated mucin, is generated by a
RT   polymorphic gene encoding splice variants with alternative amino
RT   termini.";
RL   J. Biol. Chem. 265:5573-5578(1990).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90368715; PubMed=1697589;
RA   Gendler S.J., Lancaster C.A., Taylor-Papadimitriou J., Duhig T.,
RA   Peat N., Burchell J., Pemberton L., Lalani E.-N., Wilson D.;
RT   "Molecular cloning and expression of human tumor-associated
RT   polymorphic epithelial mucin.";
RL   J. Biol. Chem. 265:15286-15293(1990).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91097524; PubMed=2268309;
RA   Lancaster C.A., Peat N., Duhig T., Wilson D.,
RA   Taylor-Papadimitriou J., Gendler S.J.;
RT   "Structure and expression of the human polymorphic epithelial mucin
RT   gene: an expressed VNTR unit.";
RL   Biochem. Biophys. Res. Commun. 173:1019-1029(1990).
RN   [5]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276413; PubMed=2351132;
RA   Wreschener D.H., Hareuveni M., Tsarfaty I., Smorodinsky N.,
RA   Horov J., Zaretsky J., Kotkes P., Weiss M., Lathe R., Dion A.,
RA   Keydar I.;
RT   "Human epithelial tumor antigen cDNA sequences. Differential splicing
RT   may generate multiple protein forms.";
RL   Eur. J. Biochem. 189:463-473(1990).
RN   [6]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast carcinoma;
RX   MEDLINE=90276414; PubMed=2112460;
RA   Hareuveni M., Tsarfaty I., Zaretsky J., Kotkes P., Horev J.,
RA   Zrihan S., Weiss M., Green S., Lathe R., Keydar I., Wreschner D.H.;
RT   "A transcribed gene, containing a variable number of tandem repeats,
RT   codes for a human epithelial tumor antigen. cDNA cloning, expression
RT   of the transfected gene and over-expression in breast cancer
RT   tissue.";
RL   Eur. J. Biochem. 189:475-486(1990).
RN   [7]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91033045; PubMed=1688329;
RA   Tsarfaty I., Hareuveni M., Horev J., Zaretsky J., Weiss M.,
RA   Jeltsch J.M., Garnier J.M., Lathe R., Keydar I., Wreschner D.H.;
RT   "Isolation and characterization of an expressed hypervariable gene
RT   coding for a breast-cancer-associated antigen.";
RL   Gene 93:313-318(1990).
RN   [8]
RP   PARTIAL SEQUENCE FROM N.A.
RX   MEDLINE=88330762; PubMed=3417635;
RA   Gendler S.J., Taylor-Papadimitriou J., Duhig T., Rothbard J.,
RA   Burchell J.;
RT   "A highly immunogenic region of a human polymorphic epithelial mucin
RT   expressed by carcinomas is made up of tandem repeats.";
RL   J. Biol. Chem. 263:12820-12823(1988).
RN   [9]
RP   SEQUENCE OF 1-169 FROM N.A.
RX   MEDLINE=90088473; PubMed=2597151;
RA   Abe M., Siddiqui J., Kufe D.;
RT   "Sequence analysis of the 5' region of the human DF3 breast
RT   carcinoma-associated antigen gene.";
RL   Biochem. Biophys. Res. Commun. 165:644-649(1989).
RN   [10]
RP   SEQUENCE OF 1-109 FROM N.A.
RC   TISSUE=Thyroid;
RX   MEDLINE=96183746; PubMed=8608966;
RA   Weiss M., Baruch A., Keydar I., Wreschner D.H.;
RT   "Preoperative diagnosis of thyroid papillary carcinoma by reverse
RT   transcriptase polymerase chain reaction of the MUC1 gene.";
RL   Int. J. Cancer 66:55-59(1996).
RN   [11]
RP   SEQUENCE OF 1-89 FROM N.A.
RC   TISSUE=Lung;
RX   MEDLINE=96181716; PubMed=8604237;
RA   Yu C.J., Yang P.C., Shew J.Y., Hong T.M., Yang S.C., Lee Y.C.,
RA   Lee L.N., Luh K.T., Wu C.W.;
RT   "Mucin mRNA expression in lung adenocarcinoma cell lines and
RT   tissues.";
RL   Oncology 53:118-126(1996).
RN   [12]
RP   SEQUENCE OF 1-46 FROM N.A.
RC   TISSUE=Breast carcinoma;
RA   Buluwela L., Liu Q., Luqmani Y.A., Gomm J.J., Coombes R.C.;
RL   Submitted (OCT-1992) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: DIRECT OR INDIRECT INTERACTION WITH ACTIN
CC       CYTOSKELETON.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN. A SECRETED FORM
CC       IS ALSO PRODUCED.
CC   -!- ALTERNATIVE PRODUCTS: VARIOUS VARIANTS ARE PRODUCED BY ALTERNATIVE
CC       SPLICING.
CC   -!- TISSUE SPECIFICITY: ABERRANTLY EXPRESSED IN HUMAN EPITHELIAL
CC       TUMORS, SUCH AS BREAST CANCER.
CC   -!- PTM: HIGHLY GLYCOSYLATED (N-AND O-LINKED CARBOHYDRATES AND SIALIC
CC       ACID).
CC   -!- POLYMORPHISM: THE NUMBER OF REPEATS IS HIGHLY POLYMORPHIC. IT
```

```
CC       VARIES FROM 21 TO 125 IN THE NORTHERN EUROPEAN POPULATION. THE
CC       MOST FREQUENT ALLELES CONTAINS 41 AND 85 REPEATS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; J05582; AAA60019.1; -.
DR   EMBL; M32738; AAA35804.1; -.
DR   EMBL; M32739; AAA35806.1; -.
DR   EMBL; J05581; AAA59876.1; -.
DR   EMBL; M61170; AAB53150.1; -.
DR   EMBL; X52229; CAA36478.1; ALT_SEQ.
DR   EMBL; X52228; CAA36477.1; ALT_SEQ.
DR   EMBL; M35093; AAB59612.1; ALT_SEQ.
DR   EMBL; Z17324; CAA78972.1; -.
DR   EMBL; Z17325; CAA78973.1; -.
DR   EMBL; M31823; AAA35757.1; -.
DR   EMBL; S81781; AAD14376.1; ALT_INIT.
DR   EMBL; S81736; AAD14369.1; ALT_INIT.
DR   EMBL; M21868; AAA59874.1; ALT_SEQ.
DR   PIR; A35175; A35175.
DR   PIR; B35175; B35175.
DR   PIR; S10218; S10218.
DR   GlycoSuiteDB; P15941; -.
DR   MIM; 158340; -.
DR   MIM; 113720; -.
DR   InterPro; IPR000082; SEA.
DR   Pfam; PF01390; SEA; 1.
DR   SMART; SM00200; SEA; 1.
DR   PROSITE; PS50024; SEA; 1.
KW   Glycoprotein; Signal; Cytoskeleton; Actin-binding; Transmembrane;
KW   Repeat; Alternative splicing.
FT   SIGNAL        1     23       POTENTIAL.
FT   CHAIN        24   1255       MUCIN 1.
FT   DOMAIN       24   1162       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM   1163   1186       POTENTIAL.
FT   DOMAIN     1187   1255       CYTOPLASMIC (POTENTIAL).
FT   DOMAIN       81    960       44 X 20 AA TANDEM REPEATS.
FT   DOMAIN     1034   1151       SEA.
FT   CARBOHYD    957    957       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    975    975       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1029   1029       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1055   1055       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1133   1133       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC     19     19       T -> TATTAPKPAT (IN ISOFORM B).
FT   VARSPLIC     20     22       MISSING (IN ISOFORM C).
FT   VARSPLIC     20     31       MISSING (IN ISOFORM D).
FT   VARSPLIC    126    905       MISSING (IN POLYMORPHIC EPITHELIAL
FT                                ISOFORM).
FT   VARSPLIC   1077   1087       FLQIYKQGGFL -> VSIGLSFPMLP (IN SECRETED
FT                                ISOFORM).
FT   VARSPLIC   1088   1255       MISSING (IN SECRETED ISOFORM).
FT   CONFLICT      2      2       T -> A (IN REF. 11).
FT   CONFLICT    134    134       P -> Q (IN REF. 9).
FT   CONFLICT    154    154       P -> Q (IN REF. 9).
FT   CONFLICT   1021   1021       S -> T (IN REF. 3).
FT   CONFLICT   1251   1251       A -> T (IN REF. 3).
SQ   SEQUENCE   1255 AA;  122072 MW;  5E28DFC4DE7D9A82 CRC64;

  Query Match             11.4%;  Score 651;  DB 1;  Length 1255;
  Best Local Similarity   27.8%;  Pred. No. 4.4e-20;
  Matches  295;  Conservative   72;  Mismatches  460;  Indels  234;  Gaps   51;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTI---K 173
          |    :||:|    |    ||                 | |     ||||     |:    |
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNK 125

Qy     174 SAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEP------APTTPKEPAPTTTKEPAPTTTK 227
           || :   ||   | ||| |   ||| :|  |      || |   ||| :|  ||   |
Db     126 PAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAHGVTSAPDT--RPAPGSTAPPAHGVT- 179

Qy     228 SAPTTPKEPAPTTP----------KKPAPTTPKEP------APTTPKEPTPTTP------ 265
           ||| |   |  | |           :||| :   |      || |   |  | |
Db     180 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 239

Qy     266 ----KEPAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPK----E 306
               ||| :  |        || |   |  |||       || |   |  | |
Db     240 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT 299

Qy     307 PAPTTTKEPSPTTPKEPAPTT---TKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP-- 361
           || |   |  | |      |:   |: || :|  ||   | ||| |   |:| :|  |
Db     300 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVT-SAPDT--RPAPGSTAPPAH 356

Qy     362 ----APTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEP 409
               || |   |  | |        || |   |  | |          || |   |  |||
Db     357 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAH 416

Qy     410 APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTT 469
             |:  :| |        |       || |   |  | |    |       | |   |  |
Db     417 GVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAPGSTA 472

Qy     470 PKA----AAPNTPKEPAPTTPK----EPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 521
           | |     :||:|   |  | |        || |   |  | |    |     :|| |  |
Db     473 PPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP----PAHGVTSAPDTRPAP 528

Qy     522 APTTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPT 569
           | |            :|||   ||       ||    :|||  | |      |:    ||
Db     529 GSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAPDTRPAPGSTAPPAHGVTS----APD 583

Qy     570 TPKEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTT 615
           |   | | |      :|| |   | | |            :|||   ||    | ||
Db     584 TRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP-PAHGVTSAP 642

Qy     616 SDKPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVS 659
             :|||  | |       :|| |   | | |            :||| :  | ||     |
Db     643 DTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS--TAPPAHGVTS 700

Qy     660 TPTTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEM 712
           | |   | :       :  :||      | |    : ::|      |   | :| |
Db     701 APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP-----DTRPAPGSTAP-- 753

Qy     713 TTTAKDKTTERDLRTTPETT-------TAAPKMTKETATTTEKTTESKITATTTQVTSTT 765
             |  |:  | | | :|       |:||   :|      |       |:  |
Db     754 --PAHGVTSAPDTRPAPGSTAPPAHGVTSAP----DTRPAPGSTAPPAHGVTSAPDTRPA 807

Qy     766 TQDTTP--FKITTLKTTTLAPKVTTTKKTITTT--EIMNKPEETAKPKDRATNSKATTPK 821
           | |    :|:  |  ||  |       |:  :    |  || |    |::  | |
Db     808 PGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 867

Qy     822 PQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPN 881
           |   : ||      |  | | |    | :|  |    :: |:  |                |
Db     868 P--GSTAPPAHGVTSAPDTRP---APGSTAPPAHGVTSAPDTRPA------------PG 909

Qy     882 QT-PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPML 940
           | | :  |   | :  | |:     |              ||   |  | : :      | :
Db     910 STAPPAHGVTSAPDTRPAPGSTAPPAH----------GVTSAPDNRPALGSTA----PPV 955

Qy     941 SDETNICNGKPVDGLTTLRNGTLVAFRGHYFWMLSPFSPPS 981
           : |:       | : |||             :||| ||
Db     956 HNVTSASGSASGSASTLVHNGTSARATTTPASKSTPFSIPS 996

RESULT   5
FP1_MYTED
ID   FP1_MYTED      STANDARD;      PRT;   875 AA.
```

```
AC   Q25460;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN (POLYPHENOLIC ADHESIVE PROTEIN) (FOOT
DE   PROTEIN 1) (MEFP1) (FRAGMENT).
GN   FP1.
OS   Mytilus edulis (Blue mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=6550;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=91025829; PubMed=1367451;
RA   Filpula D.R., Lee S.M., Link R.P., Strausberg S.L., Strausberg R.L.;
RT   "Structural and functional repetition in a marine mussel adhesive
RT   protein.";
RL   Biotechnol. Prog. 6:171-177(1990).
RN   [2]
RP   PARTIAL SEQUENCE, AND POST-TRANSLATIONAL MODIFICATIONS.
RX   MEDLINE=83135732; PubMed=6298211;
RA   Waite J.H.;
RT   "Evidence for a repeating 3,4-dihydroxyphenylalanine- and
RT   hydroxyproline-containing decapeptide in the adhesive protein of the
RT   mussel, Mytilus edulis L.";
RL   J. Biol. Chem. 258:2911-2915(1983).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: THE DECAPEPTIDE A-K-P-S-Y-P-P-T-Y-K IS POST-TRANSLATIONALLY
CC       MODIFIED AS FOLLOWING: THE SIXTH AND SEVENTH RESIDUES ARE
CC       HYDROXYLATED AND THE PENULTIMATE IS A 3,4-DIHYDROXYPHENYLALANINE
CC       (DOPA) DERIVED FROM TYROSINE.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X54422; CAA38294.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Repeat; Hydroxylation.
FT   NON_TER       1      1
FT   DOMAIN       67    870       TANDEM REPEATS OF Y-K-[PA]-K-[LP]-[ST]-Y-
FT                                 P-P-[ST].
SQ   SEQUENCE   875 AA;  100412 MW;  6EA85312748CAACE CRC64;
```

```
  Query Match             9.8%;  Score 556.5;  DB 1;  Length 875;
  Best Local Similarity   28.7%;  Pred. No. 2.1e-16;
  Matches  272;  Conservative  106;  Mismatches  375;  Indels  195;  Gaps   56;

Qy      42 TQHNKVSTSPKITTAKPINP-----------RP--SLPPNSDTSKETSLTVNKETTVET 87
           |:|  |    | : : |  |             ||  | ||    ||   | : |: :
Db       1 TKHEPVYKPVKTSYSAPYKPPTYQPLKKKVDYRPTKSYPPTYG-SKTNYLPLAKKLSSYK 59

Qy      88 KETTTTNKQTSTDG--KEKTT---SAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKG 142
              || | :|:     | | |   : |   |   |          |:   || ||: :
Db      60 PIKTTYNAKTNYPPVYKPKMTYPPTYKPKPSYPPTYKSKPTYKPKITYPPTYKAKPSY-- 117

Qy     143 PALTTPKEPTPTTPKE----PASTTPKEPTPTTIKSAPTTP-----KEPAPTTTKSAPTT 193
           |:   ||:  | | |     | :  ||   | | |  |: |     |:  |:: |: |:
Db     118 PSSYKPKKTYPPTYKPKLTYPPTYKPKPSYPPSYKTKKTYPSSYKAKPSY 177
```

```
Qy     194 PKEPAPTTTKEPA--PTTPKEPA--------PTTTKEPA-PTTTKSAPTTPK--EPAPTT 240
           |    ||   :|:  ||   :|:         ||   :|  |:| |: |: |   :  ||
Db     178 P----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTYKAKPTY 233

Qy     241 PKKPA--PTTPKEPA--PTTPKEPT--PTTPKEP---APTTKEPAPTTPKEPA--PTAPK 289
             ||:  ||   :|:  ||   :|:  ||   :|   |  | :  ||   :|:  ||
Db     234 KAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYKAKPTYKAKPSYPPTYKA 293

Qy     290 KPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA-PTTTKSAPTTTKEPA- 339
           ||:  ||   :|:  ||   :|:  ||   :||  ||   :|: | | |: ||   :|
Db     294 KPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTY 353

Qy     340 PTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPA--PTTPKEPA--PTTPKEPAP 395
           |:| |: |: |   ||   :|:     :  ||   ||:  ||   :|:  ||   :|:
Db     354 PSTYKAKPSYP----PTYKAKPSYPPTYKAKPTYKAKPSYPPTYKAKPSYPPTYKAKPSY 409

Qy     396 TTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPE 455
             | |  ||   :|       | |:|  |  |: |    |:   ||:     :  ||
Db     410 PPTYKAKPTYKAKP-------TYPST-YKAKPSYP----PSYKAKPSYPPTYKAKPTYKA 457

Qy     456 EPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKE----PAPTTPKETAPTT-- 509
           :||       |:| ||     |  ||     |  | | |      |:   ||:| | |
Db     458 KPT-------YPSTYKA----KPSYPASYKAKPSYPPTYKSKSSYPSSYKPKKTYPPTYK 506

Qy     510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTT----PKE 565
           || |   | | | |: |     ||   | | |                || : | |    |
Db     507 PKLTYKPTYK-PKPSYPPSYKPKTTYPPTYK-------------PKISYPPTYKAKPSY 551

Qy     566 PAPTTPKEPAPTTPKGT---APTTLKEPA--PTTPKKPA--PKELAPTTTKG-PTSTTSD 617
           ||     |   | | |       ||   :|:  ||   ||:   |   |:| |  |:   :
Db     552 PATYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYKAKPTYPSTYKAKPSYPPTY 611

Qy     618 KPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETPPPTTSEVSTPTTTKE---- 666
           |  |: |    ||   :|:  ||   ||  ||:|    |  ||    ::| | | |
Db     612 KAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSY 667

Qy     667 PTTIHKSPDESTPELSAEPTPKALENSPK----EPGVPTT--KTPA-----ATKPEMTTT 715
           | | :|:     |   |:|| ||   :|     :|  | |    |:       ||    |
Db     668 PPT-YKAKPSYPPTYKAKPTYKAKPTNPSTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPT 726

Qy     716 AKDKTTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKIT 775
           | | | :  | | |  | |  | :   |   | ::|        :   |    |
Db     727 YKAKPTYKAKPTYPSTYKAKP--TYKAKPTYPPTYKAK-------PSYPPTYKPKPSYPP 777

Qy     776 TLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKK---P 832
           | |: :: |     |||   |   ||: | |  :        |||  |    |   |
Db     778 TYKSKSIYPSSYKPKKTYPPT---YKPKLTYPPTYK--------PKPSYPPSYKPKITYP 826

Qy     833 TSTKKPKTMPRVRKPKTTPTP------------RKMTSTMPELNPTSR 868
           :: |   : |   | ||:  |             :  ||  |   ||:|
Db     827 STYKLKPSYPPTYKSKTSYPPTYNKKISYPSSYKAKTSYPPAYKPTNR 874
```

```
RESULT   6
ZAN_HUMAN
ID   ZAN_HUMAN      STANDARD;      PRT;  2700 AA.
AC   Q9Y493; O00218;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ZONADHESIN (FRAGMENT).
GN   ZAN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE OF 1-2379 FROM N.A.
RX   MEDLINE=99018118; PubMed=9799793;
RA   Glockner G., Scherer S., Schattevoy R., Boright A., Weber J.,
RA   Tsui L.C., Rosenthal A.;
RT   "Large-scale sequencing of two regions in human chromosome 7q22:
```

```
RT   analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci
RT   reveals 17 genes.";
RL   Genome Res. 8:1060-1073(1998).
RN   [2]
RP   SEQUENCE OF 2338-2700 FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97271566; PubMed=9126492;
RA   Gao Z., Harumi T., Garbers D.L.;
RT   "Chromosome localization of the mouse zonadhesin gene and the human
RT   zonadhesin gene (ZAN).";
RL   Genomics 41:119-122(1997).
CC   -!- FUNCTION: BINDS IN A SPECIES-SPECIFIC MANNER TO THE ZONA PELLUCIDA
CC       OF THE EGG. MAY BE INVOLVED IN GAMETE RECOGNITION AND/OR
CC       SIGNALING.
CC   -!- SUBUNIT: PROBABLY FORMS COVALENT OLIGOMERS.
CC   -!- SUBCELLULAR LOCATION: TYPE I MEMBRANE PROTEIN, EXCLUSIVELY ON THE
CC       APICAL REGION OF THE SPERM HEAD (BY SIMILARITY).
CC   -!- TISSUE SPECIFICITY: IN TESTIS; PRIMARILY IN HAPLOID SPERMATIDS.
CC   -!- DOMAIN: THE MAM DOMAINS PROBABLY MEDIATE SPERM ADHESION TO THE
CC       ZONA PELLUCIDA.
CC   -!- DOMAIN: DURING SPERM MIGRATION THROUGH THE REPRODUCTIVE TRACTS,
CC       THE MUCIN-LIKE DOMAIN MIGHT INHIBIT INAPPROPRIATE TRAPPING OF
CC       SPERMATOZOA OR PROMOTING ADHESION TO THE OVIDUCTAL ISTHMUS.
CC   -!- DOMAIN: THE VWFD DOMAINS 2 AND 3 MAY MEDIATE COVALENT
CC       OLIGOMERIZATION (BY SIMILARITY TO HUMAN INTESTINAL MUCIN MUC2).
CC   -!- SIMILARITY: CONTAINS 3.5 MAM DOMAINS.
CC   -!- SIMILARITY: CONTAINS 4.5 VWFD DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF053356; AAC78790.1; -.
DR   EMBL; U83191; AAC51208.1; -.
DR   MIM; 602372; -.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR002965; P_rich_extensn.
DR   InterPro; IPR002919; TIL.
DR   InterPro; IPR003328; TILa.
DR   InterPro; IPR001007; VWFC.
DR   InterPro; IPR001846; Vwd.
DR   Pfam; PF00629; MAM; 3.
DR   Pfam; PF01826; TIL; 4.
DR   Pfam; PF02345; TILa; 4.
DR   Pfam; PF00094; vwd; 4.
DR   PRINTS; PR01217; PRICHEXTENSN.
DR   SMART; SM00137; MAM; 2.
DR   SMART; SM00214; VWC; 1.
DR   SMART; SM00011; VWC_def; 3.
DR   SMART; SM00216; VWD; 3.
DR   PROSITE; PS01186; EGF_2; 3.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 4.
KW   Glycoprotein; Transmembrane; Cell adhesion; Repeat.
FT   NON_TER       1      1
FT   DOMAIN       <1    109       MAM 1.
FT   DOMAIN      112    136       MAM 2 (PARTIAL).
FT   DOMAIN      161    326       MAM 3.
FT   DOMAIN      322    446       MAM 4.
FT   DOMAIN      483    951       66 X HEPTAPEPTIDE REPEATS (APPROXIMATE)
FT                                (MUCIN-LIKE DOMAIN).
FT   DOMAIN      953   1065       VWFD 1 (PARTIAL).
FT   DOMAIN     1066   1454       VWFD 2.
FT   DOMAIN     1455   1861       VWFD 3.
FT   DOMAIN     1862   2292       VWFD 4.
FT   DOMAIN     2293   2684       VWFD 5.
FT   DOMAIN        ?      ?       EGF-LIKE.
FT   CARBOHYD     74     74       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1023   1023       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1099   1099       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1618   1618       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1737   1737       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1832   1832       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1878   1878       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2136   2136       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   2505   2505       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT   2374   2379       NNQKMA -> RAGPGP (IN REF. 1).
FT   NON_TER    2700   2700
SQ   SEQUENCE   2700 AA;  293013 MW;  80E60CC0B12277B1 CRC64;


  Query Match             9.7%;  Score 555.5;  DB 1;  Length 2700;
  Best Local Similarity  32.6%;  Pred. No. 5.9e-16;
  Matches  199;  Conservative   64;  Mismatches  240;  Indels  107;  Gaps   32;

Qy      58 PINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKT 117
           |:   | ||| | |      |   |||  |:  | : |              | | ||||
Db     445 PVKVLPELPPVSPVSS----TGPSETTGLTENPTISTK------------KPTVSIEKP 487

Qy     118 SAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPT-----PTTI 172
           |      ||       |  |    |: | ::| |   || | |  :   ::||     || :
Db     488 SVTTEKPT-----VPKEKPTIPTEKPTISTEK---PTIPSEKPNMPSEKPTIPSEKPTIL 539

Qy     173 KSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA-PTTTKSAPT 231
             || | |  ||     ||     |      |:|  |||| |   |  :||  ||   | ||
Db     540 TEKPTIPSE-KPTIPSEKPTISTEKPTVPTEE--PTTPTEETTTYMEEPVIPTEKPSIPT 596

Qy     232 TPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKP 291
             ::|:  | |   ||   |    : ::||      |   || |   || | | :  :|:|
Db     597 --EKPSIPTEK---PTISMEETIISTEKPTICPEKPTIPTEK---PTIPTEKSTISPEK- 647

Qy     292 APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPK 351
            |||| |  || | |    :|::|:  | | |  :  |    | |   ||     || |
Db     648 -PTTPTE-KPTIPTEKPTISTEKPTIPTEK-PTISPEKLTIPTEKLTIPT---EKPTIPT 701

Qy     352 EPSPTTTKEPAPTTPKEPAPTTPKKPA-----PTTPKEPAPTTPKEPAPTTTKKPAPTAP 406
           |    :|:|  |||| |    : :||:     || | |   |: :|   :| | |   |
Db     702 EKPTISTEE--PTTPTEETTISTEKPSIPMEKPTLPTEETTTSVEETTISTEKLTIPM-- 757

Qy     407 KEPAPTTPKETAPT-----TPKKLTPTTPEKLA-----PTTPEKPAPTTPEELAPTTPEE 456
           ::|  :| | | ||     :|:|||  | |||       || | :    : |:|  | | |
Db     758 EKPTISTEKPTIPTEKPTISPEKLTIPT-EKLTIPTEKPTIPIEETTISTEKL--TIPTE 814

Qy     457 PTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPT 516
             :||:|  :| |           || | |  || | |        ||  :| | | ||
Db     815 KPTISPEKPTISTEK----------PTIPTE-KPTIPTE-------ETTISTEKLTIPT 855

Qy     517 TLKEPAPTTPKKPAPKELAPTTTKEPTSTT------SDKPAPTTPKGTAPTTPKEPAPTT 570
             ::|  :  |   | |     :|::||  |       ::||   | |:| ||  ::
Db     856 --EKPTISPEKLTIPTEKPTISTEKPTIPTEKLTIPTEKPTIPTEKPTIPT--EKLTALR 911

Qy     571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET--A 628
           |  |:||   | |   :   ||:||       : |||   :||    | |    |:
Db     912 PPHPSPTA-TGLAALVMSPHAPSTPMTSV---ILGTTTTSRSSTGMSCP-PNARYESCAC 966

Qy     629 PTTPKEPAPT 638
           | : | | |:
Db     967 PASCKSPRPS 976


RESULT   7
EXTN_TOBAC
ID   EXTN_TOBAC     STANDARD;      PRT;   620 AA.
AC   P13983;
DT   01-JAN-1990 (Rel. 13, Created)
DT   01-JAN-1990 (Rel. 13, Last sequence update)
DT   01-MAR-1992 (Rel. 21, Last annotation update)
DE   EXTENSIN PRECURSOR (CELL WALL HYDROXYPROLINE-RICH GLYCOPROTEIN).
GN   HRGPNT3.
OS   Nicotiana tabacum (Common tobacco).
```

```
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
OC   Asteridae; euasterids I; Solanales; Solanaceae; Nicotiana.
OX   NCBI_TaxID=4097;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CV. XANTHI; TISSUE=Leaf;
RX   MEDLINE=90128263; PubMed=2612909;
RA   Keller B., Lamb C.J.;
RT   "Specific expression of a novel cell wall hydroxyproline-rich
RT   glycoprotein gene in lateral root initiation.";
RL   Genes Dev. 3:1639-1646(1989).
CC   -!- FUNCTION: HAS A SPECIALIZED STRUCTURAL FUNCTION, POSSIBLY IN
CC       THE MECHANICAL PENETRATION OF THE CORTEX AND EPIDERMIS OF THE
CC       MAIN ROOT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR MATRIX.
CC   -!- PTM: EXTENSINS CONTAIN A CHARACTERISTIC REPEAT OF THE PENTAPEPTIDE
CC       SER-PRO(4). THE PROLINE RESIDUE IS HYDROXYLATED AND THEN
CC       GLYCOSYLATED.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X13885; CAA32090.1; -.
DR   PIR; S06733; S06733.
KW   Repeat; Cell wall; Glycoprotein; Signal; Structural protein;
KW   Hydroxylation.
FT   SIGNAL        1      ?
FT   CHAIN         ?    620       EXTENSIN.
FT   REPEAT       70     73        H-A-P-P.
FT   REPEAT      148    151        H-A-P-P.
FT   DOMAIN      229    242        2 X 7 AA TANDEM REPEATS OF T-H-R-H-A-P-P.
FT   REPEAT      229    235        1.
FT   REPEAT      236    242        2.
FT   DOMAIN      205    620        CONTAINS THE SER-PRO(4) REPEATS.
FT   DOMAIN      499    600        3 X APPROXIMATE TANDEM REPEATS.
SQ   SEQUENCE   620 AA;  65406 MW;  641DD2278AB28524 CRC64;


  Query Match             9.7%;  Score 551;  DB 1;  Length 620;
  Best Local Similarity   27.2%;  Pred. No. 2.7e-16;
  Matches  170;  Conservative   64;  Mismatches  304;  Indels   88;  Gaps   24;

Qy     136 AETTTK-GPALTTP--KEPTPTT-----PKEPASTTPKE---PTPTTIKSAPTTPKEPAP 184
           || ||: |  |  |   :| |::     |  | :| |     |:|   : ||    | |
Db      24 AEATTQYGGYLPPPVTSQPPPSSIGLSPPSAPTTTPPSRGHVPSP---RHAPPRHAYPPP 80

Qy     185 TTTKSAPTTPKEP------APTTTKEPAPT---TPKEPAPTTTKEPAPTTTKSAPTTPKE 235
           :    |:    |      |:    | |:   :|  | |:    |          |:  :
Db      81 SHGHLPPSVGGPPPHRGHLPPSRGFNPPPSPVISPSHPPPSYGAPPPSHGPGHLPSHGQR 140

Qy     236 PAPTTPKKPAPT---TPKEPAPTTPKEPTPTTPKEPAPTTK--EPAPTTPKEPAPTAPKK 290
           |      |:|:    |     | |:   | : : |:| ::   | | |  :| ||
Db     141 P-------PSPSHGHAPPSGGHTPPRGQHPPSHRRPSPPSRHGHPPPPTYAQPPPTPIYS 193

Qy     291 PAPTTPKEPAPTTPKEPAPTTTKEPSPTTPK---EPAPTTTKSAPTTTKEPAPTTTKS-- 345
           |:|    :| ||    | | |  :|:|: |    :| | | : || | :   ||   |
Db     194 PSPQV--QPPPT--YSPPPPTHVQPTPSPPSRGHQPQPPTHRHAPPTHRHAPPTHQPSPL 249

Qy     346 ---APTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPAPT---TPKEPAPTTTK 399
              |:  ::| | |   | |   : | |:    | | |   | |:   :|  || : :
Db     250 RHLPPSPRRQPQPPTYSPPPPAYAQSPQPSPTYSPPPPTYSPPPPSPIYSPPPPAYSPSP 309

Qy     400 KPAPT---APKEPAPTTPKETAPTTPKKL-TPTT------PEKLAPTTPEKPAPTTPEEL 449
           | ||   :|  || : |   :|  |  |  |::      |   :|  |   :|  |  |
Db     310 PPTPTPTFSPPPPAYSPPPTYSPPPPTYLPLPSSPIYSPPPPVYSPPPPPSYSPPPPTYL 369

Qy     450 APTTPEEPTPTTPEEPAPTTPKAAAP---NTPKEPAPTTPKEPAPTTPKEPAPTTPKETA 506
           |  |  | | :   | ||  ::  |    :|  |||  |   | ||    | | |  :
Db     370 PPPPPSSPPPPSFSPPPPTYEQSPPPPPAYSPPLPAPPTYSPPPPT--YSPPPPTYAQPP 427

Qy     507 PTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEP 566
           |      | |   | |     | |    |  |  |      :| |  |  : |     |
Db     428 P------LPPTYSPPPPAYSPPPPPTYSPPPPTYSPPPPAYAQPPPPPPTYSPPPPAYSP 481

Qy     567 APTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKE 626
           | :|    |       | |   | |     | |    |   : || | .  |:| |
Db     482 PPPSPIYSPPPPQVQPLPPTFSPPPPRRIHLPPPPHRQP---RPPTPTYGQPPSPPTFSP 538

Qy     627 TAPTTPKEPAPT--TPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
           |     | |    |: | ||  : | | |     |     |   |: |   ||
Db     539 PPPRQIHSPPPPHWQPRTPTPTYGQPPSPPTFSAPPPRQIHSPPPPHRQPRPPTPTYGQP 598

Qy     685 PTPKALENSPKEPGVPTTKTPAATKP 710
           |:|             ||| :| :  |
Db     599 PSP-----------PTTYSPPSPPP 612


RESULT   8
NFH_MOUSE
ID   NFH_MOUSE      STANDARD;      PRT;  1087 AA.
AC   P19246; Q61959;
DT   01-NOV-1990 (Rel. 16, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H).
GN   NEFH OR NFH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89121513; PubMed=3220257;
RA   Julien J.-P., Cote F., Beaudet L., Sidky M., Flavell D., Grosveld F.,
RA   Mushynski W.;
RT   "Sequence and structure of the mouse gene coding for the largest
RT   neurofilament subunit.";
RL   Gene 68:307-314(1988).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=89089138; PubMed=3145094;
RA   Shneidman P.S., Carden M.J., Lees J.F., Lazzarini R.A.;
RT   "The structure of the largest murine neurofilament protein (NF-H) as
RT   revealed by cDNA and genomic sequences.";
RL   Brain Res. 464:217-231(1988).
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SWISS WEBSTER; TISSUE=Brain;
RA   Carden M.J.;
RL   Submitted (MAR-1994) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.2 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS 534
CC       TO 716 AND IS SHORTER DUE TO FRAMESHIFTS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
```

```
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------------
DR   EMBL; M24496; AAA39813.1; -.
DR   EMBL; M23349; AAA39813.1; JOINED.
DR   EMBL; M24494; AAA39813.1; JOINED.
DR   EMBL; M24495; AAA39813.1; JOINED.
DR   EMBL; M35131; AAA39809.1; ALT_FRAME.
DR   EMBL; Z31012; CAA83229.1; -.
DR   PIR; JT0368; QFMSH.
DR   PIR; A43778; A43778.
DR   MGD; MGI:97309; Nfh.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation;
KW   Repeat.
FT   DOMAIN        1     97       HEAD.
FT   DOMAIN       98    408       ROD.
FT   DOMAIN      409   1087       TAIL.
FT   DOMAIN      436    517       GLU-RICH (ACIDIC).
FT   DOMAIN      519    886       50 X 6 AA TANDEM REPEATS OF K-S-P-A-E-A.
FT   DOMAIN      887   1087       GLU/LYS-RICH.
FT   DOMAIN       98    129       COIL 1A.
FT   DOMAIN      130    141       LINKER 1.
FT   DOMAIN      142    239       COIL 1B.
FT   DOMAIN      240    261       LINKER 12.
FT   DOMAIN      262    283       COIL 2A.
FT   DOMAIN      284    287       LINKER 2.
FT   DOMAIN      288    408       COIL 2B.
FT   CONFLICT    133    133       K -> QA (IN REF. 2 AND 3).
FT   CONFLICT    199    199       A -> AR (IN REF. 2 AND 3).
FT   CONFLICT    281    281       S -> T (IN REF. 2 AND 3).
FT   CONFLICT    492    492       L -> G (IN REF. 2 AND 3).
FT   CONFLICT    551    551       P -> PREAKSP (IN REF. 3).
FT   CONFLICT    689    712       MISSING (IN REF. 3).
FT   CONFLICT    714    714       G -> A (IN REF. 3).
FT   CONFLICT    814    814       V -> M (IN REF. 2 AND 3).
FT   CONFLICT    843    843       T -> N (IN REF. 2 AND 3).
SQ   SEQUENCE   1087 AA;  116612 MW;  57BAC76A38ED1CB9 CRC64;
```

Query Match 9.4%; Score 533; DB 1; Length 1087;
Best Local Similarity 27.5%; Pred. No. 2.3e-15;
Matches 216; Conservative 81; Mismatches 330; Indels 158; Gaps 38;

```
Qy      21 DEAGSGLDNGDFKVT-----TPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKET 75
           :|   |     | :|      |  ||  | || :   |   :              : |||
Db     404 EECRIGFGPSPFSLTEGLPKIPSIST--HIKVKSEEMIKVVE-----------KSEKET 449

Qy      76 SLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPK 135
             :   :   :   |  |  :     |:|   : :  :  |:  |   :| ::  | |  :
Db     450 VIVEGQTEEIRVTEGVTEEEDKEAQGQEGEEAEEGEEKEEEELAAATSPPAEEAASPEKE 509

Qy     136 AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPK 195
           ::  |  | :  :  :|   | ||    :  :|   || |   | |    : : | :|
Db     510 TKSRVKEEAKSPGEAKSPGEAKSPA----EAKSPGEAKS-PGEAKSPGEAKSPAEPKSPA 564

Qy     196 EP-APTTTKEPAPTTPKEPAPTTTKEP----APTTTKSAPTTPKEPA----PTTPKKPA- 245
           || :|   | ||   || ||  | | |    :|:  || |   | ||    |   | ||
Db     565 EPKSPAEAKSPA--EPKSPA--TVKSPGEAKSPSEAKS-PAEAKSPAEAKSPAEAKSPAE 619

Qy     246 ---PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEPAPTTPKEPA----PTAPKK 290
             |   | ||    | | | |    :|:  | ||   :  :|   | ||    |   |
Db     620 AKSPAEAKSPAEAKSPATVKSPGEAKSPSEAKSPA---EAKSPAEAKSPAEAKSPAEVKS 676

Qy     291 P----APTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKSAPTTTKEPAPTT 342
           |    :|  || ||    | ||    :  ||   | ||    |   || |   | ||
Db     677 PGEAKSPAEPKSPAEA--KSPA----EVKSPAEAKSPAEVKSPGEAKS-PAAVKSPAEAK 729

Qy     343 TKSAPTTPKE-PSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPAPTT 397
           : :|  :| |  ||    | ||    |   | |      : | |  || | :  :  :|
Db     730 SPAAVKSPGEAKSPGEAKSPAEAKSPAEAKSPIEVKSPEKAKTPVKEGAKSPAEAKSPEK 789

Qy     398 TKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEP 457
           | |      | ||    | |  :| |      |||  |   : |   || :   | | :
Db     790 AKSPVKEDIKPPAEAKSPEKA-KSPVKEGAKPPEKAKPLDVKSPEAQTPVQEEATVPTDI 848

Qy     458 TPTTPEE-PAPTTPKAAAPNTPKEPAPTT----PKEPAPTTP-------KEPAPTTPKET 505
           |  ||:  :|   || :|   || | |:    ||:    :|        |||     :|
Db     849 RP--PEQVKSPAKEKAKSPE--KEEAKTSEKVAPKKEEVKSPVKEEVKAKEPPKKVEEEK 904

Qy     506 APTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSD--KPAPTTPKGTAPTTP 563
           |||  |  : |: ||    ||  :|   | |::|  :|::  |      |    :
Db     905 TLPTPKTEAKESKKDEAPKEAPKPKVEEKKETPTEKPKDSTAEAKKEEAGEKKKAVASEE 964

Qy     564 KEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTT 623
           : ||    || |   ||    ||  |   |  |:|:     || |: |
Db     965 ETPAKLGVKEEA--KPKEKTETTKTEAEDTKAKEPS----KPTETEKP------------ 1006

Qy     624 PKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSA 683
                 ||  |  |:|                          | :| ||  : |:|
Db    1007 --------KKEEMPAAPEK-----------------KDTKEEKTTESRKPEEK------ 1034

Qy     684 EPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETA 743
           || :|   ||      :| |  :||:     |  :|::     ||      | |:: |
Db    1035 ---PK-MEAKVKEDDKSLSKEP--SKPKTEKAEKSSSTDQKESQPPE------KTTEDKA 1082

Qy     744 TTTEK 748
           |  ||
Db    1083 TKGEK 1087
```

```
RESULT   9
TCNA_TRYCR
ID   TCNA_TRYCR     STANDARD;      PRT;  1162 AA.
AC   P23253;
DT   01-NOV-1991 (Rel. 20, Created)
DT   01-NOV-1991 (Rel. 20, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SIALIDASE (EC 3.2.1.18) (NEURAMINIDASE) (NA) (MAJOR SURFACE ANTIGEN).
GN   TCNA.
OS   Trypanosoma cruzi.
OC   Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma.
OX   NCBI_TaxID=5693;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=SILVIO X-10/4;
RX   MEDLINE=91277609; PubMed=1711561;
RA   Pereira M.E.A., Mejia J.S., Ortega-Barria E., Matzilevich D.,
RA   Prioli R.P.;
RT   "The Trypanosoma cruzi neuraminidase contains sequences similar to
RT   bacterial neuraminidases, YWTD repeats of the low density lipoprotein
RT   receptor, and type III modules of fibronectin.";
RL   J. Exp. Med. 174:179-191(1991).
RN   [2]
RP   SUBCELLULAR LOCATION.
RX   MEDLINE=91376547; PubMed=1896773;
RA   Prioli R.P., Mejia J.S., Aji T., Aikawa M., Pereira M.E.A.;
RT   "Trypanosoma cruzi: localization of neuraminidase on the surface of
RT   trypomastigotes.";
RL   Trop. Med. Parasitol. 42:146-150(1991).
CC   -!- FUNCTION: DEVELOPMENTALLY REGULATED NEURAMINIDASE IMPLICATED IN
CC       PARASITE INVASION OF CELLS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF 2,3-, 2,6- AND 2,8-GLYCOSIDIC
CC       LINKAGES JOINING TERMINAL NON-REDUCING N- OR O-ACYLNEURAMINYL
CC       RESIDUES TO GALACTOSE, N-ACETYLHEXOSAMINE, OR N- OR O-ACYLATED
CC       NEURAMINYL RESIDUES IN OLIGOSACCHARIDES, GLYCOPROTEINS,
CC       GLYCOLIPIDS OR COLOMINIC ACID.
CC   -!- SUBCELLULAR LOCATION: ATTACHED TO THE MEMBRANE BY A GPI-ANCHOR
CC       (POSSIBLE).
CC   -!- DEVELOPMENTAL STAGE: MAXIMAL ACTIVITY IN TRYPOMASTIGOTES, MINIMUM
CC       IN EPIMASTIGOTES AND NO DETECTABLE IN AMASTIGOTES.
```

```
CC   -!- MISCELLANEOUS: THE VARIABLE LENGTHS OF THE LONG TANDEM REPEAT
CC       DOMAIN COULD ACCOUNT IN PART FOR THE POLYMORPHISM OF THE TNCA
CC       PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 33 OF GLYCOSYL HYDROLASES.
CC   -!- SIMILARITY: CONTAINS 3 BNR REPEATS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M61732; AAA30255.1; -.
DR   PIR; JH0557; JH0557.
DR   HSSP; P29768; 1DIL.
DR   InterPro; IPR002860; BNR.
DR   Pfam; PF02012; BNR; 2.
KW   Hydrolase; Glycosidase; Glycoprotein; Repeat; GPI-anchor;
KW   Phosphorylation.
FT   DOMAIN        1    457       CYS-RICH.
FT   REPEAT       23     34       BNR 1.
FT   REPEAT      163    174       BNR 2.
FT   REPEAT      209    220       BNR 3.
FT   DOMAIN      458    588       FIBRONECTIN TYPE-III.
FT   DOMAIN      589   1120       44 X 12 AA TANDEM REPEATS, LTR DOMAIN.
FT   CARBOHYD    342    342       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    394    394       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1125   1125       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1162 AA;  120032 MW;  07049221897C6A40 CRC64;

  Query Match             9.3%;  Score 530.5;  DB 1;  Length 1162;
  Best Local Similarity   29.4%;  Pred. No. 3e-15;
  Matches  177;  Conservative  100;  Mismatches  212;  Indels  113;  Gaps   34;

Qy    152 TPTTPKE-PASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTP 210
          ||:|| :  | :||  |  :   || :||  || ::    |:|| : :   |    |:||
Db    599 TPSTPADSSAHSTPSTPVDS---SAHSTPSTPADSSAHGTPSTPVDSSAHGT----PSTP 651

Qy    211 KE-PAPTTTKEPAPTTTKSAPTTPKE-PAPTTPKKP-------APTTPKEPA----PTTP 257
           :  |  |   |  ::  | |:|| :  | :||  |       ||:|| : :    |:||
Db    652 ADSSAHGTPSTPVDSSAHSTPSTPVDSSAHSTPSTPVDSSAHGAPSTPADSSAHGTPSTP 711

Qy    258 KEPTP-TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPS 316
           : :   ||  || ::    |:|| : :       | :||  || ::      |     :
Db    712 VDSSAHGTPSTPADSSAHSTPSTPADSS-------AHSTPSTPADSSAHSTPSTPVDSSA 764

Qy    317 PTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA----PT 372
            ||  || ::  | |:|   || ::    |:|| : |  :|    |:|| : :    |:
Db    765 HGTPSTPADSSAHSTPST---PADSSAHGTPSTPVDSSAHST----PSTPVDSSAHGTPS 817

Qy    373 TP-KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKL----TP 427
          ||    | :||  |  ::      |     | : |  || ::    | |:||       ||
Db    818 TPVDSSAHSTPSTPVDSSAHGTPSTPVDSSAHSTPSTPADSSAHST-PSTPADSSAHGTP 876

Qy    428 TTP-EKLAPTTPEKPAPTTPEELAPTTPEE----PTPTTPEEPA----PTTP-KAAAPNT 477
          :|| :  | :||  || ::     |:|| :     ||:|| : :    |:||  ::|  |
Db    877 STPVDSSAHSTPSTPADSSAHS-TPSTPVDSSAHSTPSTPADSSAHGTPSTPVDSSAHGT 935

Qy    478 PKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPT 537
          |  || ::    | :||  || ::   | |:||  ::       | :||  |       |
Db    936 PSTPADSS----AHSTPSTPADSSAHST-PSTPADSS-------AHSTPSTPVDSSAHST 983

Qy    538 TTKEPTSTTSDKPAPTTPKGTAPTTPKE-PAPTTPKEPAPTTPKGTAPTTLKEPAPTTPK 596
                || :|  | :|     |:|| :  | :||  || ::    |  | :   | :||
Db    984 -----PSTPADSSAHST-----PSTPADSSAHSTPSTPADSSAHSTPSTPVDSSAHSTPS 1033

Qy    597 KPAPKELAPTTTKGPTSTTSDKPAPTTPK----ETAPTTPKEPAPTTPKKPAPTTPETPP 652
          |     | ::  |  || :|  | :||      :| :||  || ::    |  || ||
Db   1034 TP-----ADSSAHGTPSTPADSSAHSTPSTPVDSSAHSTPSTPADSS----AHGTPSTPA 1084

Qy    653 PTTSEVSTPTTTKEPTTIHKSPDESTP-ELSAEPTPKALENSPKEPGVPTTTKTPAATKPE 711
          :::  |||:|  : :: | :|  ||| : ||  ||                 ||| :
Db   1085 DSSAH-STPSTPAD-SSAHGTP--STPADSSAHSTP--------------STPAGSSAN 1125

Qy    712 MT 713
          |
Db   1126 GT 1127


RESULT  10
CPN_DROME
ID   CPN_DROME      STANDARD;      PRT;   865 AA.
AC   Q02910;
DT   01-OCT-1993 (Rel. 27, Created)
DT   01-OCT-1993 (Rel. 27, Last sequence update)
DT   01-FEB-1994 (Rel. 28, Last annotation update)
DE   CALPHOTIN.
GN   CPN OR CAP.
OS   Drosophila melanogaster (Fruit fly).
OC   Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta;
OC   Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
OC   Ephydroidea; Drosophilidae; Drosophila.
OX   NCBI_TaxID=7227;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165729; PubMed=8094559;
RA   Martin J.H., Benzer S., Rudnicka M., Miller C.A.;
RT   "Calphotin: a Drosophila photoreceptor cell calcium-binding protein.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1531-1535(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=CANTON-S;
RX   MEDLINE=93165730; PubMed=8434015;
RA   Ballinger D.G., Xue N., Harshman K.D.;
RT   "A Drosophila photoreceptor cell-specific protein, calphotin, binds
RT   calcium and contains a leucine zipper.";
RL   Proc. Natl. Acad. Sci. U.S.A. 90:1536-1540(1993).
CC   -!- FUNCTION: MIGHT FUNCTION AS A CALCIUM-SEQUESTERING "SPONGE" TO
CC       REGULATE THE AMOUNT OF FREE CYTOPLASMIC CALCIUM. IT BINDS 0.3 MOL
CC       OF CA+2 PER MOL OF PROTEIN.
CC   -!- SUBUNIT: HOMODIMER (PROBABLE).
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC; HYPODENSE COMPARTMENT.
CC   -!- TISSUE SPECIFICITY: SOMA AND AXONS OF PHOTORECEPTOR CELLS OF
CC       COMPOUND EYES AND OCELLI.
CC   -!- DEVELOPMENTAL STAGE: EXPRESSED EARLY IN PHOTORECEPTOR CELL
CC       DEVELOPMENT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02111; AAA28405.1; -.
DR   EMBL; L05080; AAA28420.1; -.
DR   PIR; A47282; A47282.
DR   FlyBase; FBgn0010218; Cpn.
KW   Calcium-binding.
FT   CONFLICT     36     36       A -> AVAPAVVA (IN REF. 2).
FT   CONFLICT     43     43       I -> T (IN REF. 2).
FT   CONFLICT     64     64       I -> V (IN REF. 2).
FT   CONFLICT     76     76       T -> A (IN REF. 2).
FT   CONFLICT    100    100       P -> PP (IN REF. 2).
FT   CONFLICT    126    127       VQ -> AP (IN REF. 2).
FT   CONFLICT    154    154       I -> V (IN REF. 2).
FT   CONFLICT    160    160       S -> T (IN REF. 2).
FT   CONFLICT    534    534       A -> E (IN REF. 2).
FT   CONFLICT    699    699       I -> T (IN REF. 2).
FT   CONFLICT    703    703       V -> L (IN REF. 2).
FT   CONFLICT    721    721       D -> E (IN REF. 2).
```

```
SQ   SEQUENCE   865 AA;  84781 MW;  2110417E0B0E7CFE CRC64;

  Query Match             9.3%;  Score 530;  DB 1;  Length 865;
  Best Local Similarity   26.2%;  Pred. No. 2.5e-15;
  Matches  233;  Conservative   73;  Mismatches  433;  Indels  152;  Gaps   41;

Qy    123 APTSKVLAKP-TPK---AETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTT 178
          :| |  :| | ||     |      ||   |       | | : ||  | ||     | |
Db      8 SPVSAPVAAPVTPSAVAAPVQVVSPAAVAP------APAAPIAVTPVAPPPTLASVQPAT 61

Qy    179 PKEPAPTTTKSAPTTP-KEPAPTTTKEPAPTTPKEPAPTTTKEPAPT----TTKSAPTTP 233
            |||    :|  ||     ||     | |     || :    |         |||  |
Db     62 VTIPAPAPIAAASVTPVASVAPPVVAAPTP-----PAASPVSTPVAVAQIPVAVSAPVAP 116

Qy    234 KEPAPTTPKKPAP-------TTPKEPAPTTPKEPTP-------TTPKEPAPTTKEPAPTT 279
            |  ||    |       | |   :  ||   ||        : |  || ||   |
Db    117 PVAATPTPVVQIPVAAPVIATPPVAASAPTPAAVTPVISPVIASPPVVPANTTVPVAAPV 176

Qy    280 PKEPA--PTAPKKPAPTTPKEPAPTTPKEPAPTTTKE-PSPTTPKEPAP-----TTTKSA 331
            ||  |      ||      ||   : |||    | |  | |:  ||     :   :
Db    177 AAVPAAVPVVAPVLAPAVAPAVAPVVAETPAPPPVAEIPVATIPECVAPLIPEVSVVATK 236

Qy    332 PTTTKEP---APTTTKSAPTTPKEPSPTTTKEPA-PTTPKEPAPTTPKKPAPTTPKEPAP 387
          |    ||   ||  |::    |   ||  :  ||  |    |   : : |        ||
Db    237 PLAAAEPVVVAPPATETPVVAPAAASPHVSVAPAVETAVVAPVSASTEPPVAAATLTTAP 296

Qy    388 TTP--------KEPAPTTTKKPAPTAPKEP---APTTPKET--------APTTPKKLTPT 428
            ||          : |  |     ||   ||   ||    ||          | |||  : |
Db    297 ETPALAPVVAESQVAANTVVATPPTPAPEPETIAPPVVAETPEVASVAVAETTPPVVPPV 356

Qy    429 TPEKL-APTTPEKPAPTTPEELAPTTPEEPTPTTPEEP---APTTPKAAAPNTPKEPAPT 484
            | : ||     | | |    || | |:       || |   ||:   :|   || : ||
Db    357 AAESIPAPVVATTPVPAT---LAVTDPDVTASAVPELPPVIAPSPVPSAVAETPVDLAPP 413

Qy    485 TPKEPAPTTPKEPAPTTPKETAPTTP-KGTAPTTLKE----------PAPTTPKKPAPKE 533
              |    || |     |  | ||   :|| |:              ||:     ||
Db    414 V----LPPVAAEPVPAVVAEETPETPAPASAPVTIAALDIPEVAPVIAAPSDAPAEAPSA 469

Qy    534 LAPTTTKEPTSTTSDKPAPTTPKGTAPTTP------KEPAPTTPKEP--APTTPKGTAPT 585
            ||  :  |  ||:  |  | |    || |       | |  ||  |    ||    |
Db    470 AAPIVSTPP--TTASVPETTAPPAAVPTEPIDVSVLSEAAIETPVAPPVEVTTEVAVADV 527

Qy    586 TLKEPA------PTTPKKPAPKELAPTTT----KGPTSTTSDKPAPTT----PKETAPTT 631
            | |       |  | | |  |  ||:    :   ||:|  | | |    | |   :
Db    528 APPEAAADLIIEPVEPPAPIPDLLEQTTSVPAVEAAESTSS--PIPETSLPPPNEAVASP 585

Qy    632 PKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPD----------ESTPEL 681
              || |  :| |     | |: :  : |   : |  | :: |            | ||
Db    586 EVAVAPITAPEPIP----EPEPSLATPTEPIPVEAPVVIQEAVDAVEVPVTETSTSIPET 641

Qy    682 SAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTP--ETTTAAPKMT 739
            : |  |:|:     :| :  |: |  |:         |     :: ||| |  | ||| ::: 
Db    642 TVE-FPEAVAEKVLDPAI--TEAPVTTQEPDVANINDGAPATEI-TTPAVEIVTAAAEVS 697

Qy    740 KETATTTEKTTESKITATTTQVTST------TTQDTTPFKITTLKTTTLA-PKVTTTKKT 792
              :     :|       | |       | | | :    : :  | | ::
Db    698 DIAIPVIDPPVPQEIAVAEIPETDTKPAEVIVEQSTIPIEAPVPEVSKYAEPVISEAPAA 757

Qy    793 ITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPT 852
              : |: |:       : |  |  : ||   : |  :  |     |   | ||
Db    758 EVPITAGDNPDNTSVGISEVVPTIAEKPVEEVPTS--EIPEQSSSPS--DSVPVAKITPL 813

Qy    853 PRKMTSTMPELNPTSRIAEAM---LQTTTRPNQTPNSKLVEVN-----PKS 895
            | : :|   |   :   :|:   |:     ||    :| |:      |||
Db    814 LRDLQTTDVSLLAIAATLDAIGEKLKDQKARNQQVMDRLCEIEKILGPPKS 864


RESULT  11
FP1_MYTCO
ID   FP1_MYTCO      STANDARD;      PRT;   872 AA.
AC   Q25434;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ADHESIVE PLAQUE MATRIX PROTEIN PRECURSOR (FOOT PROTEIN 1) (MCFP1).
GN   FP1.
OS   Mytilus coruscus (Sea mussel).
OC   Eukaryota; Metazoa; Mollusca; Bivalvia; Pteriomorphia; Mytiloida;
OC   Mytiloidea; Mytilidae; Mytilus.
OX   NCBI_TaxID=42192;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Foot;
RX   MEDLINE=96394686; PubMed=8798340;
RA   Inoue K., Takeuchi Y., Takeyama S., Yamaha E., Yamazaki F., Odo S.,
RA   Harayama S.;
RT   "Adhesive protein cDNA sequence of the mussel Mytilus coruscus and
RT   its evolutionary implications.";
RL   J. Mol. Evol. 43:348-356(1996).
CC   -!- FUNCTION: PROVIDES ADHESIVENESS TO THE MUSSEL'S FOOT. MUSSELS
CC       PRODUCE ONE OF THE STRONGEST WATER INSOLUBLE GLUE. THE MUSSEL'S
CC       ADHESIVE IS A BUNDLE OF THREADS, CALLED A BYSSUS, FORMED BY A
CC       FIBROUS COLLAGENOUS CORE COATED WITH ADHESIVE PROTEINS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY THE BYSSAL GLAND.
CC   -!- DOMAIN: ALMOST EXCLUSIVELY COMPOSED OF REPEATS OF A DECAPEPTIDE.
CC   -!- PTM: PROLINES IN THE REPEAT ARE HYDROXYLATED (SINGLE OR DI-) AND
CC       ALSO TYROSINE (THUS PRODUCING DOPA = 3,4-DIHYDROXYPHENYLALANINE).
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; D63777; BAA09850.1; -.
DR   InterPro; IPR002964; Adhesive_plaq.
DR   InterPro; IPR002965; P_rich_extensn.
DR   PRINTS; PR01216; ADHESIVEI.
DR   PRINTS; PR01217; PRICHEXTENSN.
KW   Signal; Repeat; Hydroxylation.
FT   SIGNAL        1     20       POTENTIAL.
FT   CHAIN        21    872       ADHESIVE PLAQUE MATRIX PROTEIN.
FT   DOMAIN       21     41       NONREPETIVE LINKER.
FT   DOMAIN      124    872       TANDEM REPEATS OF Y-K-[PS]-K-[IP]-[ST]-Y-
FT                                P-[PST]-[ST].
FT   DOMAIN      184    192       NONAPEPTIDE 1.
FT   DOMAIN      213    221       NONAPEPTIDE 2.
SQ   SEQUENCE   872 AA;  101677 MW;  98CC70D7C75FF3C4 CRC64;

  Query Match             9.1%;  Score 518.5;  DB 1;  Length 872;
  Best Local Similarity   29.2%;  Pred. No. 7.3e-15;
  Matches  276;  Conservative   80;  Mismatches  407;  Indels  181;  Gaps   54;

Qy     19 VVDEAGSGLDNGDFKVTTPDTST--TQH-------NKVSTSPKITTAKPINPRPSLPPNS 69
          |   | ||   | :| | | :     ::|          ||: | | |  :|   | | |
Db     29 VYGSAYSGASAGAYK-TLPGSHPYGSKHVPVYKPMNKIPT-PYI--SKKSYPAPYKPKGY 84

Qy     70 DTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVL 129
            :|    |   :|           | :|    : |  | : ::    | |       |:
Db     85 YPTKRYQPTYGSKTNYPPIYKPIAKKLSSYKAIKTTYPAYKAKTSYPPSYK-----HKIT 139

Qy    130 AKPTPKAETTTKGPALTTPKEPT-PTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTK 188
            || |       | :| |  ||    |  | |  ||   | | |     ||   | | |
Db    140 YPPTYK-------PKITYP--PTYKQKPSYPPSYKPKTTYPPTYK-----PKITYPPTYK 185

Qy    189 SAPT-TPKEPAPT--TTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTT--PKK 243
            |: || :|  |   | :|  | |     ||  ::|: |     |  ||   | |  ||
Db    186 RKPSYTPYKPKATYPPTYKPKITYP----PTYKRKPSYT-----PYKPKTTYPPTYKPKI 236

Qy    244 PAPTTPKEPAPTT----PKEPTPTT--PKEPAPTTKEPAPTTPKEPAPT-APKKPAPTTP 296
            |:  |  |        |:  | |  ||   | | :| |: |    ||  ||   | |
```

```
Db    237 SYPSIYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYP----PTYKPKVTYPPTY 292

Qy    297 KEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPT-----TTKEPAPTTTKSAPTTPK 351
          | | |: |     |  |  |:   ||   ||  |  |:       :|   ||: ||  | |
Db    293 K-PKPSYPPTYKPKITYPPT-YKPKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYPP 350

Qy    352 EPSPTTTKEPA-PTTPKEPAPTTPKKP-APT-TPK-------EPAPTTPKEPAPTTTKKP 401
          |  |  |      |  |    |||  :||  ||        :| |: |     | ||  |
Db    351 TYKPKITYPPTYKPKPSYPPSYKPKKTYSPTYKPKITYPPTYKPKPSYPPSYKPKTTYPP 410

Qy    402 APTAPKEPAPTTPKETAP-----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEE 456
          ||   | | |  |        : |   ||   |::      || |: |    |
Db    411 T-YKPKISYPPTYKPKASYVSSYKSKKTYPPTYKPKISYPPTYKPKPSYPPTYKPKITYP 469

Qy    457 PT-PTTPEEPAPTTPKAAAPNTPK-EPAPTTPKEPAPTTP-----KEPAPTT--PKETAP 507
          ||    |  |    ||   | | | :|:  || :  |: |      |   ||:   |:| |
Db    470 PTYKPKPSYPPTYKPKITYPPTYKRKPSYPTPYKQKPSYPPIYKSKSSYPTSYKSKKTYP 529

Qy    508 TT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPT-TPKGTAPTTPK 564
           |  || | | | |       ||    |    | ||  ||      :  ||  || : | | |
Db    530 PTYKPKITYPPTYK------PKPSYPPSYKPKTTYPPTYKPKIRYPPTYKPKASYPPTYK 583

Qy    565 EPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP----KKPAPKELAPTTTKGPTSTTSDKPA 620
                    || | | | | | |: |    :||:   :  : :  ||:  | |
Db    584 --------------PKITYPPTYK-PKPSYPTPYKQKPSYPPIYKSKSSYPTAYKSKKTY 628

Qy    621 PTT--PKETAPTTPKEPAPTTPK--KPAPTTPET-------PPPTTSEVSTPTTTKEPTT 669
          | |  || | | | | | | |: |   :|  | | |        |     |: | | : :   |
Db    629 PPTYKPKITYPPTYK-PKPSYPPSYRPKITYPPTYKPKKSYPQAYKSKGSYPPSYQPKKT 687

Qy    670 IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKT----TERDL 725
            |    |: :  || |     ||    || |    |||    : | ||     | :
Db    688 Y---PPSYKPKKTYPPTYK-----PKISYPPTYK----TKPSYPASYKRKTSYPPTYKPK 735

Qy    726 RTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPK 785
           : | |  | |     :   | | |  :  ::   ::    |   |   :: |     ||
Db    736 ISYPSTYKAKP-----SYPPTYKPKPSYASSYKPKIRYPPTYKPKPSYASSYK-----PK 785

Qy    786 VTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ---KPTKAPKKP-TSTKKPK-T 840
          :           :  | |||     | |:: ||:    ||  ||    |: ||| |
Db    786 I-------------RYPPTYKPKP----SYASSYKPKIRYPPTYKPKPSYASSYKPKIT 827

Qy    841 MPRVRKPKTT--PTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQ 882
           |   ||| :  || :   :  |   |       |     : |:|
Db    828 YPPTYKPKISYPPTYKPKITYPPTYKPKISYPPAYKPKISYPSQ 871


RESULT  12
MUC1_XENLA
ID   MUC1_XENLA     STANDARD;      PRT;   662 AA.
AC   Q05049;
DT   01-OCT-1994 (Rel. 30, Created)
DT   01-OCT-1994 (Rel. 30, Last sequence update)
DT   01-OCT-1994 (Rel. 30, Last annotation update)
DE   INTEGUMENTARY MUCIN C.1 (FIM-C.1) (FRAGMENT).
OS   Xenopus laevis (African clawed frog).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus.
OX   NCBI_TaxID=8355;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Skin;
RX   MEDLINE=93077556; PubMed=1447205;
RA   Hauser F., Hoffmann W.;
RT   "P-domains as shuffled cysteine-rich modules in integumentary mucin
RT   C.1 (FIM-C.1) from Xenopus laevis. Polydispersity and genetic
RT   polymorphism.";
RL   J. Biol. Chem. 267:24620-24624(1992).
CC   -!- FUNCTION: COULD BE INVOLVED IN DEFENSE AGAINST MICROBIAL
CC       INFECTIONS. PROTECTS THE EPITHELIA FROM EXTERNAL ENVIRONMENT.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- ALTERNATIVE PRODUCTS: A NUMBER OF DIFFERENT FORMS OF THE PROTEIN
CC       MAY BE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: SKIN.
CC   -!- PTM: EXTENSIVELY O-GLYCOSYLATED.
CC   -!- SIMILARITY: CONTAINS 6 P-TYPE (TREFOIL) DOMAINS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; L02115; AAA74725.1; -.
DR   PIR; A45155; A45155.
DR   HSSP; P04002; 1WFA.
DR   InterPro; IPR000519; P_trefoil.
DR   Pfam; PF00088; trefoil; 6.
DR   SMART; SM00018; P; 6.
DR   PROSITE; PS00025; P_TREFOIL; 6.
KW   Repeat; Amphibian skin; Glycoprotein; Alternative splicing.
FT   NON_TER       1      1
FT   DOMAIN       81    144       8 X 8 AA APPROXIMATE TANDEM REPEATS,
FT                                ALA/THR-RICH.
FT   REPEAT       81     88       1-1.
FT   REPEAT       89     96       1-2.
FT   REPEAT       97    104       1-3.
FT   REPEAT      105    112       1-4.
FT   REPEAT      113    120       1-5.
FT   REPEAT      121    128       1-6.
FT   REPEAT      129    136       1-7.
FT   REPEAT      137    144       1-8.
FT   DOMAIN      161    202       P-TYPE 1.
FT   DOMAIN      218    301       8 X APPROXIMATE TANDEM REPEATS, THR-RICH.
FT   REPEAT      218    224       2-1.
FT   REPEAT      225    239       2-2.
FT   REPEAT      240    249       2-3.
FT   REPEAT      250    259       2-4.
FT   REPEAT      260    275       2-5.
FT   REPEAT      276    287       2-6.
FT   REPEAT      288    294       2-7.
FT   REPEAT      295    301       2-8.
FT   DOMAIN      306    347       P-TYPE 2.
FT   DOMAIN      353    394       P-TYPE 3.
FT   DOMAIN      402    522       12 X APPROXIMATE TANDEM REPEATS,
FT                                THR-RICH.
FT   REPEAT      402    411       3-1.
FT   REPEAT      412    419       3-2.
FT   REPEAT      420    431       3-3.
FT   REPEAT      432    443       3-4.
FT   REPEAT      444    453       3-5.
FT   REPEAT      454    460       3-6.
FT   REPEAT      461    472       3-7.
FT   REPEAT      473    479       3-8.
FT   REPEAT      480    491       3-9.
FT   REPEAT      492    498       3-10.
FT   REPEAT      499    515       3-11.
FT   REPEAT      516    522       3-12.
FT   DOMAIN      525    566       P-TYPE 4.
FT   DOMAIN      572    613       P-TYPE 5.
FT   DOMAIN      620    661       P-TYPE 6.
FT   DISULFID    162    188       BY SIMILARITY.
FT   DISULFID    172    187       BY SIMILARITY.
FT   DISULFID    182    199       BY SIMILARITY.
FT   DISULFID    307    333       BY SIMILARITY.
FT   DISULFID    317    332       BY SIMILARITY.
FT   DISULFID    327    344       BY SIMILARITY.
FT   DISULFID    354    380       BY SIMILARITY.
FT   DISULFID    364    379       BY SIMILARITY.
FT   DISULFID    374    391       BY SIMILARITY.
FT   DISULFID    526    552       BY SIMILARITY.
FT   DISULFID    536    551       BY SIMILARITY.
FT   DISULFID    546    563       BY SIMILARITY.
```

```
FT   DISULFID    573    599       BY SIMILARITY.
FT   DISULFID    583    598       BY SIMILARITY.
FT   DISULFID    593    610       BY SIMILARITY.
FT   DISULFID    621    647       BY SIMILARITY.
FT   DISULFID    631    646       BY SIMILARITY.
FT   DISULFID    641    658       BY SIMILARITY.
FT   VARIANT     276    276       K -> E.
FT   VARIANT     354    354       C -> R.
FT   VARIANT     415    415       T -> A.
SQ   SEQUENCE   662 AA;  67774 MW;  F085277F1ED2FD40 CRC64;
```

Query Match 8.9%; Score 509.5; DB 1; Length 662;
Best Local Similarity 28.1%; Pred. No. 1.4e-14;
Matches 223; Conservative 48; Mismatches 276; Indels 247; Gaps 30;

```
Qy      92 TTNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP 151
           ||     :  ||: | :|: : : |||:|       :| | ||  |     |   ||
Db       3 TTAAAVAATGKDTTAAAEGSAAAEKTAA-----AGEVSAPPT--AAVAATGEDATT---- 51

Qy     152 TPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPK 211
             |: |    |     ||||   || |    ||||    ||||    ||||
Db      52 --------AAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPTTAT 103

Qy     212 EPAPTTTKEPAPTTTKS-APTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPA- 269
             || |  | |||  | ||||      :|    ||||    | :  :  : :: :|
Db     104 GKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKERSTSSSSEEEHCH 163

Qy     270 --PTTKEPAPT---TPKEPAPTAPKKPAPTTPK----------EPAPTTPKEPAPTTTKE 314
             |: :|   :   | |:      ||     ||          :|   : :|   ||||
Db     164 VKPSKREMCGSKGITKKQ----CKKKNCCFDPKGHGGIHCFHRKPKGHSHEEHTTTTTK- 218

Qy     315 PSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP 374
               |||| : | |||    ||||            ||| :  |||
Db     219 ---------APTTIQIATTTT---TPTTT-----------TTTTKATPTT--------- 245

Qy     375 KKPAPTTPKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLA 434
                      ||  :  ||||           || | |  ||    ||||
Db     246 ------------TTTTKATPTTT-----------TTTKATTTTT----TPTT----- 270

Qy     435 PTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTP 494
                  ||    | |||    | |||
Db     271 ---------TTTTTKATTTPTTTTTTTP------------------------------- 289

Qy     495 KEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTT 554
                  |  ||   |  ||       | |          :|    :   :
Db     290 ----------TTTTTKATTTTTTTSGECKMEPSK----------REDCGYSGITESQCR 328

Qy     555 PKG-----TAPTT-----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELA 604
            ||     : | |               ||:     |    |    |    :|    : :
Db     329 TKGCCFDSSIPQTKWCFYTLSQVADCKVEPSQRVDCGFRGIT----ADQCRQKNCCFDSS 384

Qy     605 PTTTKGPTSTTSDKPAPTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTT 664
             : ||    :||   |  |   | |||   |  ||  |   |||      |:  :|||||
Db     385 ISGTKWCFYSTSQVAA--TKTTTTPTTTTTPTTTTTTKATTTTP------TTTTTTPTTT 436

Qy     665 KEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERD 724
             ||                | ||              || ||  | |  |||   | |
Db     437 TTTTT----------------TTKA----------TTTTPTTTTPTTTTT---KAT--- 463

Qy     725 LRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAP 784
            |||  |||  |     | |||: |      |||  |:|||  ||  | ||    ||
Db     464 -TTTPTTTTTTP-----TTTTTKAT-------TTTPTTTTTTPTTTTTKATTTTPTTTTT 510

Qy     785 KVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQKPTKAPKKPTSTKKPK----T 840
            |||| | |||    | |    |  ||        : |  :|     :|  : |    :
Db     511 TTTTTKATTTTTSGECKME----PSKRADCGYPGITESQCRSKGCCFDSSIPQTKWCFYS 566

Qy     841 MPRVRKPKTTPTPR 854
           :|:|    |  |: |
Db     567 LPQVADCKVAPSSR 580
```

```
RESULT  13
RPB1_HUMAN
ID   RPB1_HUMAN     STANDARD;      PRT;  1970 AA.
AC   P24928;
DT   01-MAR-1992 (Rel. 21, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95347616; PubMed=7622068;
RA   Mita K., Tsuji H., Morimyo M., Takahashi E., Nenoi M.,
RA   Ichimura S., Yamauchi M., Hongo E., Hayashi A.;
RT   "The human gene encoding the largest subunit of RNA polymerase II.";
RL   Gene 159:285-286(1995).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X63564; CAA45125.1; -.
DR   EMBL; X74874; CAA52862.1; -.
DR   EMBL; X74873; CAA52862.1; JOINED.
DR   EMBL; X74872; CAA52862.1; JOINED.
DR   EMBL; X74871; CAA52862.1; JOINED.
DR   EMBL; X74870; CAA52862.1; JOINED.
DR   PIR; S21054; S21054.
DR   MIM; 180660; -.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 43.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1067   1067       W -> L (IN REF. 2).
FT   CONFLICT   1449   1449       D -> Y (IN REF. 2).
SQ   SEQUENCE   1970 AA;  217205 MW;  6876FC25692A657E CRC64;
```

Query Match 8.8%; Score 503.5; DB 1; Length 1970;

```
  Best Local Similarity   33.3%;  Pred. No. 5.8e-14;
  Matches  181;  Conservative  101;  Mismatches  167;  Indels   95;  Gaps   47;

Qy     151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
           |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
           |:|   :||:   :||:|   :||:|   :||:|   :||:|            |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
            :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
           :||:|   :||:    :||:|   :||:|  | ||:|   :|.|:|   :||:|
Db    1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
           ||:|   :||:|   :||:|   :||:| : :|:::|:     ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828

Qy     501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
           :|   :||:|| |          ||:|   |:  |  ||: |   | || | :||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPTSPK--YSPTSPKYSPTSPK-YS 1875

Qy     560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
           ||:|   :|||||  :||:|  |   |      ||:||       :||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
           :||:|   :||:||    :||:|   :||:|   .  |  :::|           |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     677 STPE 680
           |  |
Db    1966 SDEE 1969


RESULT  14
RPB1_MOUSE
ID   RPB1_MOUSE     STANDARD;      PRT;  1970 AA.
AC   P08775;
DT   01-NOV-1988 (Rel. 09, Created)
DT   01-MAR-1992 (Rel. 21, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   DNA-DIRECTED RNA POLYMERASE II LARGEST SUBUNIT (EC 2.7.7.6) (RPB1).
GN   POLR2A OR RPO2-1 OR RPII215.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=87280135; PubMed=3038894;
RA   Ahearn J.M. Jr., Bartolomei M.S., West M.L., Cisek L.J., Corden J.L.;
RT   "Cloning and sequence analysis of the mouse genomic locus encoding
RT   the largest subunit of RNA polymerase II.";
RL   J. Biol. Chem. 262:10695-10705(1987).
RN   [2]
RP   SEQUENCE OF 1587-1970 FROM N.A.
RX   MEDLINE=86068017; PubMed=2999785;
RA   Corden J.L., Cadena D.L., Ahearn J.M. Jr., Dahmus M.E.;
RT   "A unique structure at the carboxyl terminus of the largest subunit
RT   of eukaryotic RNA polymerase II.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:7934-7938(1985).
RN   [3]
RP   REVISIONS, AND PRESENCE OF AN ADDITIONAL EXON.
RX   MEDLINE=92178992; PubMed=1542581;
RA   Wintzerith M., Acker J., Vicaire S., Vigneron M., Kedinger C.;
RT   "Complete sequence of the human RNA polymerase II largest subunit.";
RL   Nucleic Acids Res. 20:910-910(1992).
CC   -!- FUNCTION: DNA-DEPENDENT RNA POLYMERASE CATALYZES THE TRANSCRIPTION
CC       OF DNA INTO RNA USING THE FOUR RIBONUCLEOSIDE TRIPHOSPHATES AS
CC       SUBSTRATES.
CC   -!- CATALYTIC ACTIVITY: N NUCLEOSIDE TRIPHOSPHATE = N PYROPHOSPHATE +
CC       RNA(N).
CC   -!- SUBUNIT: RNA POLYMERASE II CONSISTS OF 12 DIFFERENT SUBUNITS.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   -!- PTM: THE TANDEM 7 RESIDUES REPEATS CAN BE HIGHLY PHOSPHORYLATED.
CC       THE PHOSPHORYLATION ACTIVATES POL2.
CC   -!- MISCELLANEOUS: THREE DISTINCT ZINC-CONTAINING RNA POLYMERASES ARE
CC       FOUND IN EUKARYOTIC NUCLEI: POLYMERASE I FOR THE RIBOSOMAL RNA
CC       PRECURSOR, POLYMERASE II FOR THE MRNA PRECURSOR, AND POLYMERASE
CC       III FOR 5S AND TRNA GENES.
CC   -!- SIMILARITY: BELONGS TO THE RNA POLYMERASE BETA' CHAIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M12130; AAA40071.1; -.
DR   EMBL; M14101; AAA40071.1; JOINED.
DR   PIR; A28490; A28490.
DR   MGD; MGI:98086; Rpo2-1.
DR   InterPro; IPR000684; RNA_polII_repeat.
DR   InterPro; IPR000722; RNA_pol_A.
DR   InterPro; IPR002879; RNA_pol_A2.
DR   Pfam; PF00623; RNA_pol_A; 1.
DR   Pfam; PF01854; RNA_pol_A2; 1.
DR   PROSITE; PS00115; RNA_POL_II_REPEAT; 42.
KW   Transferase; DNA-directed RNA polymerase; Transcription; Zinc; Repeat;
KW   DNA-binding; Nuclear protein; Phosphorylation; Zinc-finger.
FT   ZN_FING      71     87       C2H2-TYPE (POTENTIAL).
FT   DOMAIN     1590   1958       CARBOXYL-TERMINAL 7-RESIDUE REPEATS.
FT   CONFLICT   1498   1498       P -> R (IN REF. 1 AND 2).
FT   CONFLICT   1499   1536       MISSING (IN REF. 1 AND 2).
SQ   SEQUENCE   1970 AA;  217175 MW;  7D76F38FD92A657E CRC64;


  Query Match             8.7%;  Score 498.5;  DB 1;  Length 1970;
  Best Local Similarity   33.1%;  Pred. No. 9.2e-14;
  Matches  180;  Conservative  101;  Mismatches  168;  Indels   95;  Gaps   47;

Qy     151 PTPTTPKEPASTTPKE-PTPTTIKSAPTTPKEPAPTTTKSAPTTPKE-----PAPTTTKE 204
           |:|     || |   :  ||     :|:      |     :|:   :     |  :
Db    1507 PSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFSPSAASDASGFSPGYSPAWS 1566

Qy     205 PAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEP--APTTPKEPTP 262
           | | :|  | |::   |:|    |   :|  ||   |: |   ||: |  :||:|   :|
Db    1567 PTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPA-YEPRSPGGYTPQSPSYSPTSPSY-SP 1624

Qy     263 TTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKE 322
           |:|   :||:   :||:|   :||:|   :||:|   :||:|            |||:|
Db    1625 TSPSY-SPTSPNYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY---------SPTSPSY 1671

Qy     323 PAPTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTP 382
            :||:   :||:    :||:   :||:|   |||:    :||:|   :||:|   :||:|
Db    1672 -SPTSPSYSPTSPSY-SPTSPSYSPTSPSY-SPTSPSY-SPTSPSY-SPTSPSY-SPTSP 1725

Qy     383 KEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPA 442
           :||:|   :||:    :||:|   :||:|  | ||:|   :||:|   :||:|
Db    1726 SY-SPTSPSY-SPTSPSY-SPTSPNY-SPTSPNYT-PTSP-SYSPTSP-SYSPTSPNY-T 1777

Qy     443 PTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEP--APT 500
           ||:|   :||:|   :||:|   :||:| : :|:::|:     ||: |   ||  |  :|:
Db    1778 PTSP-NYSPTSPSY-SPTSPSY-SPTSP-SYSPSSPR----YTPQSPT-YTPSSPSYSPS 1828
```

```
Qy     501 TPKETAPTTPKGTAPTTLKEPAPTTPK-KPAPKELAPTTTKEPTSTTSDKPAPTTPKGTA 559
           :|   :||:|| |         ||:|   |:  |  | : |   | || | :||:||  :
Db    1829 SP-SYSPTSPKYT---------PTSPSYSPSSPEYTPASPK--YSPTSPKYSPTSPK-YS 1875

Qy     560 PTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPT-STTSDK 618
           ||:|   :|||||  :||:|  |    |      ||:||       :|||:   || | || |
Db    1876 PTSPTY-SPTTPKY-SPTSP--TYSPTSPVYTPTSPK------YSPTS---PTYSPTSPK 1922

Qy     619 PAPTTPKETAPTTPKEP--APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676
           :|||:|   :|||:||    :|||:|  :|||:|      |  :::|          |||:
Db    1923 YSPTSP-TYSPTSPKGSTYSPTSPGY-SPTSP------TYSLTSPAI---------SPDD 1965

Qy     677 STPE 680
           |  |
Db    1966 SDEE 1969
```

```
RESULT  15
NFH_RAT
ID   NFH_RAT        STANDARD;      PRT;   831 AA.
AC   P16884; Q63368;
DT   01-AUG-1990 (Rel. 15, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   NEUROFILAMENT TRIPLET H PROTEIN (200 KDA NEUROFILAMENT PROTEIN)
DE   (NEUROFILAMENT HEAVY POLYPEPTIDE) (NF-H) (FRAGMENT).
GN   NEFH OR NFH.
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Brain;
RX   MEDLINE=89065087; PubMed=3143606;
RA   Breen K.C., Robinson P.A., Wion D., Anderton B.H.;
RT   "Partial sequence of the rat heavy neurofilament polypeptide (NF-H).
RT   Identification of putative phosphorylation sites.";
RL   FEBS Lett. 241:213-218(1988).
RN   [2]
RP   SEQUENCE OF 37-831 FROM N.A.
RX   MEDLINE=88309090; PubMed=2457365;
RA   Dautigny A., Pham-Dinh D., Roussel C., Felix J.M., Nussbaum J.L.,
RA   Jolles P.;
RT   "The large neurofilament subunit (NF-H) of the rat: cDNA cloning and
RT   in situ detection.";
RL   Biochem. Biophys. Res. Commun. 154:1099-1106(1988).
RN   [3]
RP   SEQUENCE OF 1-89 AND 243-313 FROM N.A.
RX   MEDLINE=87080760; PubMed=2878828;
RA   Robinson P.A., Wion D., Anderton B.H.;
RT   "Isolation of a cDNA for the rat heavy neurofilament polypeptide
RT   (NF-H).";
RL   FEBS Lett. 209:203-205(1986).
RN   [4]
RP   SEQUENCE OF 318-831 FROM N.A.
RX   MEDLINE=89184647; PubMed=2928342;
RA   Lieberburg I., Spinner N., Snyder S., Anderson J., Goldgaber D.,
RA   Smulowitz M., Carroll Z., Emanuel B.S., Breitner J., Rubin L.;
RT   "Cloning of a cDNA encoding the rat high molecular weight
RT   neurofilament peptide (NF-H): developmental and tissue expression in
RT   the rat, and mapping of its human homologue to chromosomes 1 and
RT   22.";
RL   Proc. Natl. Acad. Sci. U.S.A. 86:2463-2467(1989).
CC   -!- FUNCTION: NEUROFILAMENTS USUALLY CONTAIN THREE IF PROTEINS: L, M,
CC       AND H WHICH ARE INVOLVED IN THE MAINTENANCE OF NEURONAL CALIBER.
CC       NF-H HAS AN IMPORTANT FUNCTION IN MATURE AXONS THAT IS NOT
CC       SUBSERVED BY THE TWO SMALLER NF PROTEINS.
CC   -!- PTM: THERE ARE A NUMBER OF REPEATS OF THE TRIPEPTIDE K-S-P, NFH IS
CC       PHOSPHORYLATED ON A NUMBER OF THE SERINES IN THIS MOTIF. IT IS
CC       THOUGHT THAT PHOSPHORYLATION OF NFH RESULTS IN THE FORMATION OF
CC       INTERFILAMENT CROSS BRIDGES THAT ARE IMPORTANT IN THE MAINTENANCE
CC       OF AXONAL CALIBER.
CC   -!- PTM: PHOSPHORYLATION SEEMS TO PLAY A MAJOR ROLE IN THE FUNCTIONING
CC       OF THE LARGER NEUROFILAMENT POLYPEPTIDES (NF-M AND NF-H), THE
CC       LEVELS OF PHOSPHORYLATION BEING ALTERED DEVELOPMENTALLY AND
CC       COINCIDENT WITH A CHANGE IN THE NEUROFILAMENT FUNCTION.
CC   -!- SIMILARITY: BELONGS TO THE INTERMEDIATE FILAMENT FAMILY.
CC   -!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN FROM POSITION 783
CC       ONWARD AND IS LONGER DUE TO A FRAMESHIFT.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M37227; AAA41693.1; ALT_FRAME.
DR   EMBL; X13804; CAA32038.1; ALT_FRAME.
DR   EMBL; M21964; AAA41695.1; -.
DR   EMBL; J04517; AAA41692.1; -.
DR   PIR; A30796; A30796.
DR   PIR; A25649; A25649.
DR   PIR; B25649; B25649.
DR   PIR; S02003; S02003.
DR   InterPro; IPR001664; IF.
DR   Pfam; PF00038; filament; 1.
DR   PROSITE; PS00226; IF; 1.
KW   Intermediate filament; Coiled coil; Neurone; Phosphorylation; Repeat.
FT   NON_TER       1      1
FT   DOMAIN      276    641       51 X 3 AA TANDEM REPEATS OF K-S-P.
FT   CONFLICT    164    164       L -> I (IN REF. 2).
FT   CONFLICT    185    185       I -> S (IN REF. 2).
FT   CONFLICT    193    193       L -> T (IN REF. 2).
FT   CONFLICT    199    199       M -> T (IN REF. 2).
FT   CONFLICT    346    346       K -> N (IN REF. 1).
FT   CONFLICT    373    373       A -> V (IN REF. 1 AND 4).
FT   CONFLICT    482    482       G -> E (IN REF. 2 AND 4).
FT   CONFLICT    485    485       P -> S (IN REF. 2).
FT   CONFLICT    570    571       RK -> KE (IN REF. 2 AND 4).
FT   CONFLICT    591    591       P -> T (IN REF. 2 AND 4).
FT   CONFLICT    727    727       A -> V (IN REF. 4).
FT   CONFLICT    757    759       AAP -> GST (IN REF. 4).
FT   CONFLICT    769    769       T -> L (IN REF. 2).
FT   CONFLICT    775    775       R -> P (IN REF. 2 AND 4).
SQ   SEQUENCE   831 AA;  89486 MW;  1B0973C3F13EF768 CRC64;
```

```
  Query Match             8.7%;  Score 497;  DB 1;  Length 831;
  Best Local Similarity   27.6%;  Pred. No. 5.2e-14;
  Matches  195;  Conservative   59;  Mismatches  307;  Indels  146;  Gaps   35;

Qy      37 PDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQ 96
           |  ||  | || :  ||    :                 : ||| :   :   ::  |  | :
Db     197 PSMST--HIKVKSEEKIKVVE-----------KSEKETVIVEEQTEEIQVTEEVTEEED 242

Qy      97 TSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEP----T 152
              |:|:  : :  :    ||     | ::  | |    |  || |     | |    :
Db     243 KEAQGEEEEEAEEGGEEAATTS-----PPAEEAASP----EKETKSPVKEEAKSPAEAKS 293

Qy     153 PTTPKEPA-STTPKEPTPTTIKSAPTTPKEPA----PTTTKSAPTTPKEPA----PTTTK 203
           |   | || : :| |     :  :|   | ||    |   || |   | ||    | | |
Db     294 PAEAKSPAEAKSPAEVKSPAVAKSPAEVKSPAEVKSPAEAKS-PAEAKSPAEVKSPATVK 352

Qy     204 EP----APTTPKEPA----PTTTKEPA----PTTTKSAPTTPKEPA----PTTPKKPA-- 245
           |     :|   | ||     |   | ||     | : || |   | ||     |   | ||
Db     353 SPGEAKSPAEAKSPAEVKSPAEAKSPAEAKSPASVKS-PGEAKSPAEAKSPAEVKSPATV 411

Qy     246 --PTTPKEPA----PTTPKEP----TPTTPKEPAPTTKEP----APTTPKEPA----PTA 287
           |   | ||     | | | |    :|   | || : | |    :|   | ||    |
Db     412 KSPVEAKSPAEVKSPVTVKSPAEAKSPVEVKSPA-SVKSPSEAKSPAGAKSPAEAKSPVV 470

Qy     288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPA----PTTTKS-----APTTTKEP 338
```

```
Db    471 AKSPAEA--KSPAGAKPPAEAKSPAEAKSPAEAKSPAEAKSPAEAKSPVEVKSPEKAKSP 528

Qy    339 APTTTKS-----APTTPKEPSPTTTKEPA----PTTPKEPAPTTPKKPAPTTPKEPAPTT 389

Db    529 VKEGAKSLAEAKSPEKAKSPVKEEIKPPAEVKSPEKAKSPMRKEAKSPEKAKTLDVKSPE 588

Qy    390 PKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEEL 449

Db    589 AKPPAKEEAKRPADIRSPEQVKSPAKEEAKSPEKEETRT--EKVAPKKEEVKSPV--EEV 644

Qy    450 APTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTT 509

Db    645 ---KAKEPPKKVEEEKTPATPKTEVKESKKDEAPKEAQKP-KAEEKEPLTEKPKDSPGEA 700

Qy    510 PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPT 569

Db    701 KKEEA----KEKKAAAPEEETPAKLGVKEEAKPKEKAEDAKA-----------KEPSKP 744

Qy    570 TPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAP 629

Db    745 SEKE----KPK-------KEEVPAAPEKKDTKE---------EKTTESKKREEKPKMEAK 784

Qy    630 TTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDE 676

Db    785 AKEED-------KGLPQEPSKPKTEKAEKSSSTDQKDSQPSEKAPED 824
```

Search completed: April 26, 2002, 16:19:50
Job time: 375 sec

GenCore version 4.5
Copyright (c) 1993 - 2000 Compugen Ltd.

OM protein - protein search, using sw model

Run on: April 26, 2002, 16:17:05 ; Search time 49.78 Seconds
(without alignments)
480.986 Million cell updates/sec

Title: US-09-556-246-1_COPY_200_1263
Perfect score: 5698
Sequence: 1 VKDNKKNRTKKKPTPKPPVV.....VAALSTAKYKNWPESVYFFK 1064

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 212252 seqs, 22503292 residues

Total number of hits satisfying chosen parameters: 212252

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : Issued_Patents_AA:*
1: /cgn2_6/ptodata/2/iaa/5A_COMB.pep:*
2: /cgn2_6/ptodata/2/iaa/5B_COMB.pep:*
3: /cgn2_6/ptodata/2/iaa/6A_COMB.pep:*
4: /cgn2_6/ptodata/2/iaa/6B_COMB.pep:*
5: /cgn2_6/ptodata/2/iaa/PCTUS_COMB.pep:*
6: /cgn2_6/ptodata/2/iaa/backfiles1.pep:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	% Query Match	Length	DB	ID	Description
1	522	9.2	805	4	US-09-103-429A-4	Sequence 4, Appli
2	508	8.9	1837	3	US-08-928-361B-5	Sequence 5, Appli
3	506.5	8.9	744	6	5202236-25	Patent No. 5202236
4	498.5	8.7	786	4	US-09-103-429A-3	Sequence 3, Appli
5	489	8.6	1721	3	US-08-700-651-5	Sequence 5, Appli
6	489	8.6	1721	3	US-08-928-361B-6	Sequence 6, Appli
7	488.5	8.6	826	1	US-07-638-431-2	Sequence 2, Appli
8	488.5	8.6	826	5	PCT-US92-00018-2	Sequence 2, Appli
9	476.5	8.4	652	6	5202236-13	Patent No. 5202236
10	452	7.9	960	4	US-09-219-849-5	Sequence 5, Appli
11	424.5	7.4	1867	2	US-08-479-537A-5	Sequence 5, Appli
12	424.5	7.4	1867	4	US-09-083-116-5	Sequence 5, Appli
13	424.5	7.4	2035	2	US-08-479-537A-2	Sequence 2, Appli
14	424.5	7.4	2035	4	US-09-083-116-2	Sequence 2, Appli
15	419.5	7.4	2476	2	US-08-276-967-2	Sequence 2, Appli
16	419	7.4	1185	4	US-09-041-886-23	Sequence 23, Appl
17	417	7.3	829	1	US-08-642-255-132	Sequence 132, App
18	417	7.3	829	1	US-08-397-633A-53	Sequence 53, Appl
19	417	7.3	837	1	US-08-175-155-68	Sequence 68, Appl
20	417	7.3	837	1	US-08-477-509B-103	Sequence 103, App
21	417	7.3	837	1	US-08-642-255-101	Sequence 101, App
22	417	7.3	837	2	US-08-707-237A-75	Sequence 75, Appl
23	417	7.3	837	3	US-08-482-085B-103	Sequence 103, App
24	417	7.3	897	1	US-08-397-633A-50	Sequence 50, Appl
25	413.5	7.3	907	3	US-08-783-774-2	Sequence 2, Appli
26	413.5	7.3	907	5	PCT-US95-04611A-19	Sequence 19, Appl
27	404.5	7.1	408	1	US-07-609-716-65	Sequence 65, Appl
28	404.5	7.1	408	4	US-08-475-411A-65	Sequence 65, Appl
29	404.5	7.1	408	4	US-08-478-029A-65	Sequence 65, Appl
30	398.5	7.0	682	1	US-08-642-255-126	Sequence 126, App
31	398.5	7.0	682	1	US-08-397-633A-36	Sequence 36, Appl
32	381	6.7	1848	4	US-08-296-791-6	Sequence 6, Appli
33	381	6.7	1848	5	PCT-US95-10661A-6	Sequence 6, Appli
34	376	6.6	960	4	US-09-219-849-6	Sequence 6, Appli
35	374.5	6.6	1537	1	US-08-325-267A-2	Sequence 2, Appli
36	373	6.5	1231	3	US-08-904-263A-4	Sequence 4, Appli
37	371	6.5	761	2	US-08-707-237A-84	Sequence 84, Appl
38	371	6.5	762	1	US-08-642-255-114	Sequence 114, App
39	371	6.5	762	1	US-08-397-633A-26	Sequence 26, Appl
40	370.5	6.5	1064	1	US-08-642-255-62	Sequence 62, Appl
41	368	6.5	762	1	US-08-642-255-120	Sequence 120, App
42	368	6.5	762	1	US-08-397-633A-31	Sequence 31, Appl
43	367.5	6.4	1187	1	US-08-320-559-28	Sequence 28, Appl
44	367.5	6.4	1187	3	US-08-545-860D-28	Sequence 28, Appl
45	367.5	6.4	1187	5	PCT-US94-04496-28	Sequence 28, Appl

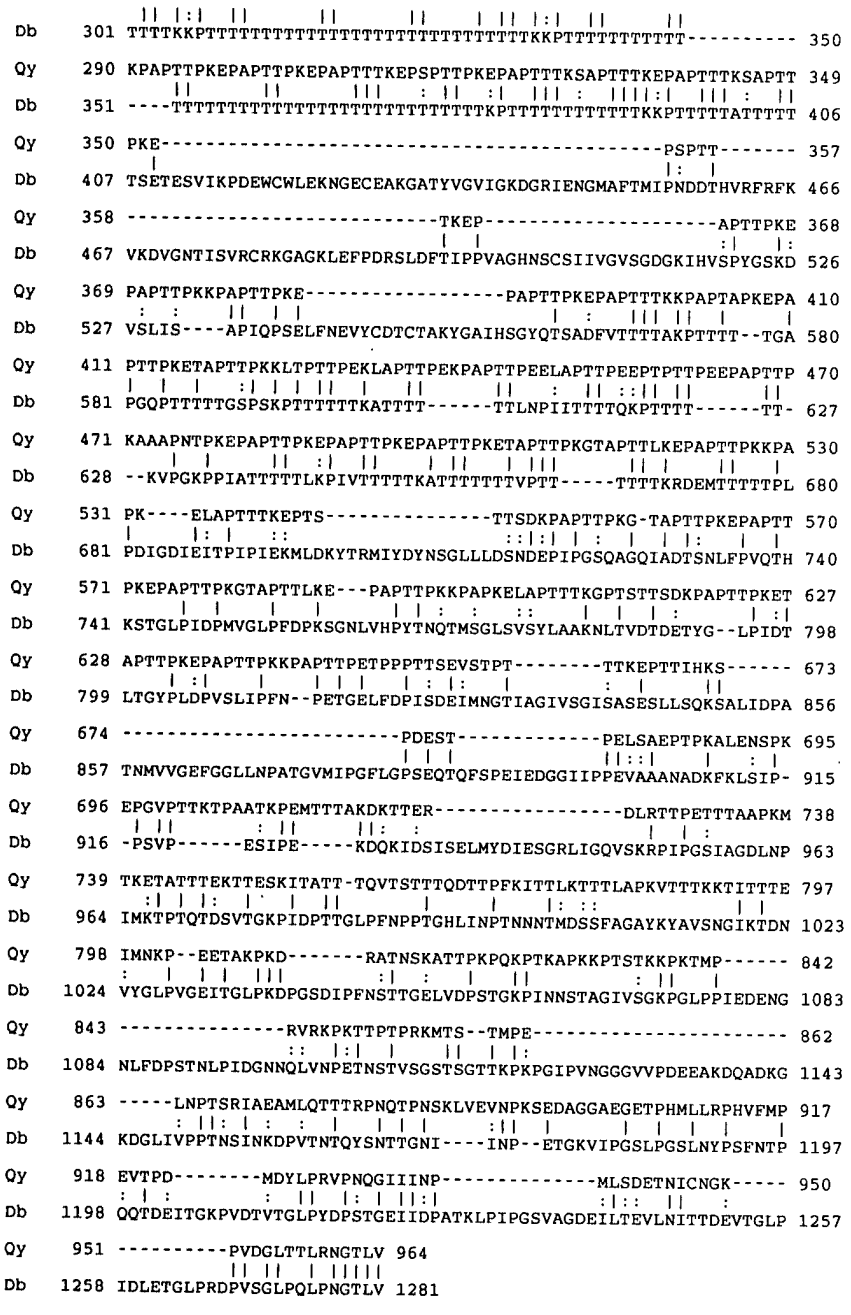
ALIGNMENTS

```
RESULT 1
US-09-103-429A-4
; Sequence 4, Application US/09103429A
; Patent No. 6187558
;  GENERAL INFORMATION:
;    APPLICANT:  Granados, Robert R
;    APPLICANT:  Wang, Ping
;    TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;    TITLE OF INVENTION:  cDNA and Related Products and Methods
;    NUMBER OF SEQUENCES:  4
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;      STREET:  118 No. 6187558th Tioga
;      CITY:  Ithaca
;      STATE:  NY
;      COUNTRY:  USA
;      ZIP:  14850
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/09/103,429A
;      FILING DATE:  24-JUN-1998
;      CLASSIFICATION:  800
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Michaels, Christopher A
;      REGISTRATION NUMBER:  34,390
;      REFERENCE/DOCKET NUMBER:  BTI-39
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (607) 256-2000
;      TELEFAX:  (607) 256-3628
;  INFORMATION FOR SEQ ID NO:  4:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  805 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
;    HYPOTHETICAL:  NO
;    ANTI-SENSE:  NO
;    ORIGINAL SOURCE:
;      ORGANISM:  Trichoplusia ni
;      TISSUE TYPE:  peritrophic membrane
US-09-103-429A-4

  Query Match             9.2%;  Score 522;  DB 4;  Length 805;
  Best Local Similarity  27.8%;  Pred. No. 7e-27;
```

```
  Matches  226;  Conservative   33;  Mismatches  211;  Indels  342;  Gaps   40;

Qy    125 TSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAP-----TTP 179
          |: |    |:     | |||  |    |  | |     | |  |             |
Db      9 TALGLVAARPEVSDAEKNPALHEPHPDXP--PAEQXXLLPXEYDCTKFYYCEYGLKFIAP 66

Qy    180 KEPAPTT-----------TKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTK 227
          :: || |           :  | |  || ||    || | :  |||||:  |||||
Db     67 RDCAPGTEFKFSAQTCVHAALAGCTLPGPPAETT---QAPATTQ--APTTTQ--APTTTT 119

Qy    228 SAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTA 287
           ||||  :   ||           |||| : || |     | ||: |  ||
Db    120 QAPTTTTQATTTT----------QAPTTTQAPTTT----QATTTQAPTTTT-------- 156

Qy    288 PKKPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTTTKEPAPTTTKSAP 347
              ||||  : ||||  : |||||:          |||||: |||||:  ||||   |
Db    157 ---QAPTTTTQ-APTTTTQ-APTTTQ----------APTTTQ-APTTTQ--APTTITQAA 198

Qy    348 TTPKEPSPTTTKEPAPTTP--KEPAPTTPKKPAPTTPKEPAPTTPK-------------- 391
          |||   :|  |  || |||    || |||  ||||:    ||  |
Db    199 TTPAATTPAAT-TPAATTPAATTPAATTPGVPAPTS----APVWPPICELLPNGCPADFD 253

Qy    392 ------------------------------------------------------EPAPTTTK 399
                                                               |||  |:
Db    254 IHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVECDGEISPAPPVTE 313

Qy    400 K---------------PA---------------------------------------- 402
                          ||
Db    314 GNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERRCGAGTHFSFELQ 373

Qy    403 --------------------------------PTAP--KEPAPT-----------TPKET 417
                                           || |   || |             | |:
Db    374 QCDHIELVGCTLPGGESEEVDVDEDACTGWYCPTEPIEWEPLPNGCPADFSIDHLLPHES 433

Qy    418 ------------------------APTTPKKLTPTT----------PEKLAPTTPEKPAP 443
                                  :| |    :| |           :   |     ||
Db    434 DCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSDNQCTSTAAPTAAP 493

Qy    444 TTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPK 503
          |    |||      |:|   || | |  |||  |    ||    ||||     ||||
Db    494 TAAPTAAPTAAPTAAPSTVVPPA-TPPATAAPVPPTTAIPT----PAPTAAPTAAPTTAA 548

Qy    504 ETAPTT---PKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAP 560
            :|||   |  |||    ||||    |  |:  | |  ||:  :   |||      ||
Db    549 PESPTTVTVPPTAAPTA----APTT----AVPEIPITVTSAPTAAPT--AAPTAAPTAAP 598

Qy    561 TTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTST-TSDKP 619
          ||     |||      |:|   ||||   ||| |           | |  ||:  |:  |
Db    599 TTAVPEIPTT-----VTSPPTAAPTT-AAPAPNT----------TVTVPPTAAPTTAAP 641

Qy    620 APTTPKETAPTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTP 679
          ||     |  | |   ||||   ||| |  | ||| :  : |     ||  |       |
Db    642 AP----NTTVTAPPTAAPTT-AAPAPNTTVTVPPTAAPTAAP-----PTVAH------AP 685

Qy    680 ELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
            :| |              | ||  || | ||
Db    686 NTTAAP------------VTTTSAPATT-PE 703


RESULT   2
US-08-928-361B-5
; Sequence 5, Application US/08928361B
; Patent No. 6071518
;  GENERAL INFORMATION:
;    APPLICANT:  Petersen, Carolyn
;    TITLE OF INVENTION:  PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
;    TITLE OF INVENTION:  THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
;    TITLE OF INVENTION:  FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
;    TITLE OF INVENTION:  SPECIES INFECTIONS
;    NUMBER OF SEQUENCES:  30
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  PETERS, VERNY, JONES & BIKSA
;      STREET:  385 Sherman Avenue, Suite 6
;      CITY:  Palo Alto
;      STATE:  CA
;      COUNTRY:  USA
;      ZIP:  94306-1840
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/928,361B
;      FILING DATE:  12-SEP-1997
;      CLASSIFICATION:
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 60/026,062
;      FILING DATE:  13-SEP-1996
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Verny, Hana
;      REGISTRATION NUMBER:  30,518
;      REFERENCE/DOCKET NUMBER:  480.76-1(HV)
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  650-324-1677
;      TELEFAX:  650-324-1678
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1837 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
US-08-928-361B-5


  Query Match             8.9%;  Score 508;  DB 3;  Length 1837;
  Best Local Similarity   21.3%;  Pred. No. 1.5e-25;
  Matches  295;  Conservative  100;  Mismatches  494;  Indels  496;  Gaps   48;

Qy     33 KVTTPDTSTTQHNKVSTSPKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTT 92
          | ||  |:||     :|:   ||         :    : |:  |: |    ||  |  |||
Db    307 KHTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 366

Qy     93 TNKQTSTDGKEKTTSAKETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPT 152
          |   |:|  |: ||:   | :   |:       |:      |   |||     || |  |
Db    367 TTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTKKPTT 426

Qy    153 PTTPKEPASTTPKEPTPTTIKSAPTTPKEP--APTTTKSAPTTPKEPAPTTTKEPAPTTP 210
           ||      :||    | ||  :  || |:|    ||| :  ||      |||     ||
Db    427 TTTTTTTTTTTTTTTTTTTTTTTTTTKKPTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 486

Qy    211 KEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKEPAPTTPKE----------- 259
           | ||||:|  ||| :  || |:|  ||      || :  :  | |
Db    487 TTTATTTTKKPTTTTTTTTTTTTKKPTTTTTATTTTTTSETESVIKPDEWCWLEKNGECE 546

Qy    260 ------------------------PTPTT--------------------------PK 266
                                     | |                          |
Db    547 AKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFKVKDVGNTISVRCGKGAGKLEFPD 606

Qy    267 EPAPTTKEP--------------------APTTPKEPA-PTAPKKP------------- 291
              | |                      :|   |: :  :|| :|
Db    607 RSLDFTIPPVAGHNSCSIIVGVSGGGKIHVSPYGSKDVSLISAPIQPCELFNEVYCDTCT 666

Qy    292 ---------------APTTPKEPAPTT---PKEPAPTTTKEPS-PTTPKEPAPTTTKS- 330
                          ||  :|  ||    | :|  |||  || |||      ||| :
Db    667 AKYGAIHSGYQTSADFVTTTTAKPTTTTGAPGQPTTTTTGSPSKPTTTTTTKATTTTTI 726

Qy    331 ----APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTPKKPAPTTPKEPA 386
               |||::|  |||    |  |   : ||| :|  ||     | ||      |||
Db    727 LNPIITTTTQKPTTTTTTKVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT----- 781

Qy    387 PTTPKEPAPTTTKKPAP-------------------------------TAPKEPAP--- 411
            || |    |||  | |                               :  || |
Db    782 TTTTKRDEMTTTTTPLPDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQ 841
```

```
Qy    412 ------------TTPKETA-PTTP------------------------------KKL 425
Db    842 AGQIADTSNLFPVQTHKSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNL 901

Qy    426 T-----------------PTTPEKLAPTTPE---------------------------- 439
Db    902 TVDTDETYGLPIDTLTGYPLDPVSLIPFNPETGELFDPISDEIMNGTIAGIVSGISASES 961

Qy    440 ----KPAPTTPEE----------LAPTT---------PEEPTPTTP--EEPAPTTPKAAA 474
Db    962 LLSQKSAPIDPATNMVVGEFGGLLNPATGVMIPGSLGPSEQTPFSPEIEDGGIIPPEVAA 1021

Qy    475 PNTPKEPAPTTPKEPAPTTPKE---------------------------------PA 498
Db   1022 ANADKFKLSIPPSVPESIPEKDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNPI 1081

Qy    499 PTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGT 558
Db   1082 MKTPTQTDSVTGKPIDPTT---GLPFNP--PTGHLINPTNNNTMDSSFAGAYKYAVSNGI 1136

Qy    559 APTT----PKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTP-----------KKPAPK-- 601
Db   1137 KTDNVYGLPVDEITGLPKDPVSDIPFNSTTGELVDPSTGKPINNYTAGIVSGKRGLPPIE 1196

Qy    602 ----ELAPTTTKGP-----------TSTT-------SDKPAPTTPKETAPTTPKEPA--- 636
Db   1197 DENGNLFDPSTKLPIDGNNQLVNPETNSTVSGSTSGSTKPKPGIPVNGGGVVPDEEAKDQ 1256

Qy    637 -----------PTTPKKPAPTTPETPPPTTSEVSTPTTTKE-PTTIHKS---PDESTPEL 681
Db   1257 ADKGKDGLIVPPTNSINKDPVTNTQYSNTTGNIINPETGKVIPGSLPGSLNYPSFNTPQQ 1316

Qy    682 SAEPTPKALE---NSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKM 738
Db   1317 TDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP-I 1375

Qy    739 TKETATTTEKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE- 797
Db   1376 DLETGLPRDP------VSGLPQLPNGTLVD-----------------PSNKKPIPGSHS 1411

Qy    798 --IMNKPEETAKPKDRATNSKATTPKPQKP-TKAP--KKPTSTKKPKTMPRVRKPKT--- 849
Db   1412 GFINGTSGEQSHEKDPSTG------KPLDPNTGLPFDEDSGSLINPETGDKLQGSHSGTF 1465

Qy    850 TPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPK-SEDA---------- 898
Db   1466 MPVPGKPQGENGGIMTPEQILEAL-------NKLPTSNEVNISPRPSSDAVPDRPTNTWW 1518

Qy    899 ---------------GGAEGETPHMLL------RPHVFMPEVTPDMDYLPRVPNQGIII 936
Db   1519 NKISGQTFQVDGKKTIPGSAASVIHTALGTPTQTDPTTGLP--SDPSTGLPFIPGFNVLV 1576

Qy    937 NPMLSD-------------ETNIC-----------NGKPVDGLTTL---RNGTLV-AF 966
Db   1577 DPQTGEQIKGSVPYVSLYVKEKNIVTEAAYGLPVDPKTGFPIDPISYLPFAKNGELIDPI 1636

Qy    967 RGHYF 971
Db   1637 SGKYF 1641


RESULT   3
5202236-25
;Patent No. 5202236
;     APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;     TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;     NUMBER OF SEQUENCES: 39
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER: US/07/528,762
;       FILING DATE: 25-MAY-1990
;       APPLICATION NUMBER: 82,456
;       FILING DATE: 07-AUG-1987
;       APPLICATION NUMBER: 933,945
;       FILING DATE: 24-NOV-1986
;       APPLICATION NUMBER: 650,128
;       FILING DATE: 13-SEP-1984
;SEQ ID NO:25:
;       LENGTH: 744
5202236-25


  Query Match             8.9%;  Score 506.5;  DB 6;  Length 744;
  Best Local Similarity   29.9%;  Pred. No. 6.7e-26;
  Matches  239;  Conservative   78;  Mismatches  308;  Indels  175;  Gaps   50;

Qy     51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
Db     24 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 53

Qy    111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
Db     54 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 92

Qy    167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
Db     93 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 151

Qy    224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEP--- 275
Db    152 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPKKTYPPTYKPKLT 209

Qy    276 -APTTPKEPA--PTAPKKPA--PTTPKEPA--PTTPKEPA--PTTTKEPS--PTTPKEPA 324
Db    210 YPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPS 269

Qy    325 -PTTTKSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKKP 377
Db    270 YPSTYKAKPSYT---YPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAKP 322

Qy    378 A--PTTPKEPA--PTTPKEPA--PTTTKKPA--PTAPKEPAPTTPKETAPTTPKKLTPTT 429
Db    323 SYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKVKPTYKAKPTYPS 382

Qy    430 PEKLAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKA--AAPNT----P 478
Db    383 TYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYKAKPTYPSTYKAKP 438

Qy    479 KEPAPTTPKEPAPTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPK 532
Db    439 SYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPSTYKAKSSYPPTYKAKP 498

Qy    533 ELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA- 591
Db    499 SYPPTYKAKPTYPSTYKAKPTYK--ARPSYP----PTYKAKPS------YPPTYKAKPSY 546

Qy    592 -PTTPKKPAPKELAPTTTKGPT--STTSDKPAPTTPKETAPTTPKEPAPTTPKKPA--PT 646
Db    547 PPTYKAKPTYK-AKPTYKAKPTYPSTYKAKPSYSPTYKAKPSYP----PTYKAKPSYPPT 601

Qy    647 ---TPETPPPTTSEVSTPTTTKE----PTTIHKSPD-----ESTPELSAEPT-PKALENS 693
Db    602 YKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAK 661

Qy    694 PKEPGVPTTKTPAA------TKPEMTTTAKDKTTERDLRTTPETTTAAP--------KMT 739
Db    662 PSYP--PTYKAKPSYPPTYKAKPSYPPTYKAKPTYKAKPTYPSTYKAKPSYPPTHKAKPT 719

Qy    740 KETATTTEKTTESKITATTT 759
Db    720 YKAKPTYPSTYKAKPTYPST 739


RESULT   4
US-09-103-429A-3
```

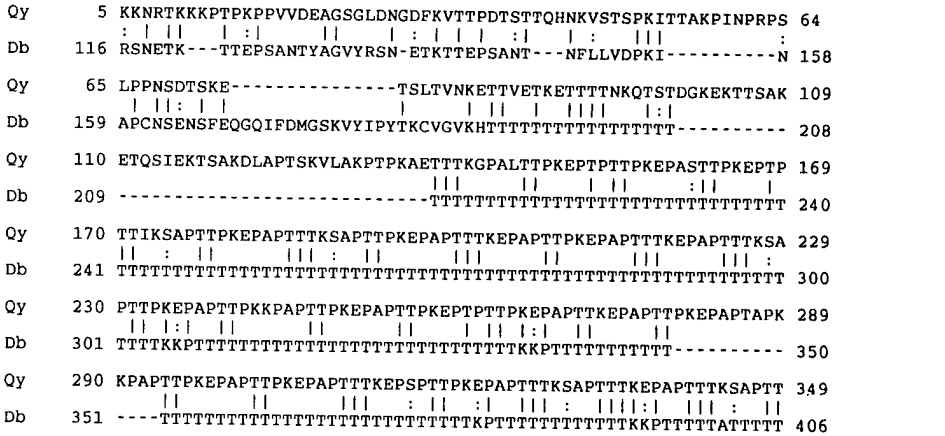
```
; Sequence 3, Application US/09103429A
; Patent No. 6187558
; GENERAL INFORMATION:
;   APPLICANT:  Granados, Robert R
;   APPLICANT:  Wang, Ping
;   TITLE OF INVENTION:  A No. 6187558el Invertebrate Intestinal Mucin
;   TITLE OF INVENTION:  cDNA and Related Products and Methods
;   NUMBER OF SEQUENCES:  4
;   CORRESPONDENCE ADDRESS:
;     ADDRESSEE:  Brown, Pinnisi & Michaels, P.C.
;     STREET:  118 No. 6187558th Tioga
;     CITY:  Ithaca
;     STATE:  NY
;     COUNTRY:  USA
;     ZIP:  14850
;   COMPUTER READABLE FORM:
;     MEDIUM TYPE:  Floppy disk
;     COMPUTER:  IBM PC compatible
;     OPERATING SYSTEM:  PC-DOS/MS-DOS
;     SOFTWARE:  PatentIn Release #1.0, Version #1.30
;   CURRENT APPLICATION DATA:
;     APPLICATION NUMBER:  US/09/103,429A
;     FILING DATE:  24-JUN-1998
;     CLASSIFICATION:  800
;   ATTORNEY/AGENT INFORMATION:
;     NAME:  Michaels, Christopher A
;     REGISTRATION NUMBER:  34,390
;     REFERENCE/DOCKET NUMBER:  BTI-39
;   TELECOMMUNICATION INFORMATION:
;     TELEPHONE:  (607) 256-2000
;     TELEFAX:  (607) 256-3628
; INFORMATION FOR SEQ ID NO:  3:
;   SEQUENCE CHARACTERISTICS:
;     LENGTH:  786 amino acids
;     TYPE:  amino acid
;     STRANDEDNESS:  single
;     TOPOLOGY:  linear
;   MOLECULE TYPE:  protein
;   HYPOTHETICAL:  NO
;   ANTI-SENSE:  NO
;   ORIGINAL SOURCE:
;     ORGANISM:  Trichoplusia ni
;     TISSUE TYPE:  peritrophic membrane
US-09-103-429A-3


  Query Match             8.7%;  Score 498.5;  DB 4;  Length 786;
  Best Local Similarity   29.1%;  Pred. No. 2.4e-25;
  Matches  200;  Conservative   25;  Mismatches  207;  Indels  255;  Gaps   32;

Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAPTTTKSAP 191
          | | || ||: || |      ||| : | :||          ||||  :   |||  ||
Db     93 PGPPAE-TTQAPATT----QAPTTTQAPTTTT----------QAPTTTTQ--ATTTTQAP 135

Qy    192 TTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTPKKPAPTTPKE 251
          || :  |||||:  | ||    ||||| : |||||  ||||  : |||| :  |||| :
Db    136 TTTQ--APTTTQ--ATTT---QAPTTTTQ-APTTTTQAPTTTTQ-APTTTQ--APTTTQ- 183

Qy    252 PAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPKKPAPTTPKEPAPTT-PKEPA-- 308
           |||| : ||  |      |    || |||    | |   || |||  ||||: |  |
Db    184 -APTTTQAPTTITQAATTPAATTPAATTPAATTPAA-TTPAATTPGVPAPTSAPVWPPIC 241

Qy    309 ----------------------------------------------PTTTKEPSPTTPK- 321
                                                         |   :  ||   :
Db    242 ELLPNGCPADFDIHLLIPHDKYCNLFYQCSNGYTFEQRCPEGLYFNPYVQRCDSPANVEC 301

Qy    322 ----EPAPTTT---------------------------------------------- 328
              |||  |
Db    302 DGEISPAPPVTEGNEDEDIDIGDLLDNGCPANFEIDWLLPHGNRCDKYYQCVHGNLVERR 361

Qy    329 --------------------------KSAPTTTKEPAPTTTKSAPTTPKEPSPTTTKEP 361
                                     :|      | |  |    || | |  |     |
Db    362 CGAGTHFSFELQQCDHIELVGCTLPGGESEEVDVDEDA-CTGWYCPTEPIEWEPLPNGCP 420

Qy    362 APTT-----PKEP----------------------APTTPKKPAPTT---------- 381
          |  :     | |                      :| |    :| |
Db    421 ADFSIDHLLPHESDCGQYLQCVHGQTIARPCPGNLHFSPATQSCESPVTAGCQVFECDSD 480

Qy    382 ---PKEPAPTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTP 438
               |||     |||     ||||   |:   |  | | |   : |||     ||
Db    481 NQCTSTAAPTAAPTAAPTAAPTAAPTA--APSTVVPPATPPATAAPVPPTT---AIPT-- 533

Qy    439 EKPAPTTPEELAPTTPEEPTPTTPEEPAPTTPKAAAPNT--PKEPAPTTPKEPAPTTPKE 496
            ||||     ||||    :|||   | ||    ||| |  |: |   |    |||
Db    534 --PAPTAAPTAAPTTAAPESPTTVTVP-PTAAPTAAPTTAVPEIPITVT---SAPTAAPT 587

Qy    497 PAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPK 556
            |||     ||||     |||:  |    |   ||   || ||                | |
Db    588 AAPTAAPTAAPTTAVPEIPTTVTSPPTAAPTTAAP---APNTT-----------VTVPP 632

Qy    557 GTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTS 616
            ||||   ||| |     ||     || |:                 || ||  | :|||
Db    633 TAAPTT-AAPAPNTTVTVPPTAAPTAAPPTVAH-------------APNTTAAPVTTTS 677

Qy    617 DKPAPTTPKETAPTTPKEPAPTTPKKP 643
            || |  |     |  | |  |  |
Db    678 ---APATTPEDDDIDP--PLPNDPINP 699


RESULT   5
US-08-700-651-5
; Sequence 5, Application US/08700651B
; Patent No. 6015882
; GENERAL INFORMATION:
;  APPLICANT: PETERSEN, CAROLYN
;  APPLICANT: LEECH, JAMES
;  APPLICANT: NELSON, RICHARD, C.
;  APPLICANT: GUT, JIRI
;  TITLE OF INVENTION: VACCINES, ANTIBODIES, PROTEINS, GLYCOPROTEINS, DNAS AND RNAS
;  TITLE OF INVENTION: FOR PROPHYLAXIS AND TREATMENT OF Cryptosporidium parvum
;  TITLE OF INVENTION: INFECTIONS
;  FILE REFERENCE: 480.19-4(HV)
;  CURRENT APPLICATION NUMBER: US/08/700,651B
;  CURRENT FILING DATE: 1997-08-14
;  EARLIER APPLICATION NUMBER: 08/415,751
;  EARLIER FILING DATE: 1995-04-03
;  NUMBER OF SEQ ID NOS: 15
;  SOFTWARE: PatentIn Ver. 2.0
; SEQ ID NO 5
;   LENGTH: 1721
;   TYPE: PRT
;   ORGANISM: Cryptosporidium parvum
US-08-700-651-5


  Query Match             8.6%;  Score 489;  DB 3;  Length 1721;
  Best Local Similarity   22.1%;  Pred. No. 2.4e-24;
  Matches  284;  Conservative   82;  Mismatches  476;  Indels  442;  Gaps   41;

Qy      5 KKNRTKKKPTPKPPVVDEAGSGLDNGDFKVTTPDTSTTQHNKVSTSPKITTAKPINPRPS 64
          : | ||   | :|     ||    | : | | |  :|   |  :   |||          :
Db    116 RSNETK---TTEPSANTYAGVYRSN-ETKTTEPSANT---NFLLVDPKI----------N 158

Qy     65 LPPNSDTSKE---------------TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAK 109
          | ||: | |                |     | ||  |  ||||   |:|
Db    159 APCNSENSFEQGQIFDMGSKVYIPYTKCVGVKHTTTTTTTTTTTTTTTTTT---------- 208

Qy    110 ETQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTP 169
                                    |||     ||     | ||     :||     |
Db    209 ----------------------------TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 240

Qy    170 TTIKSAPTTPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPAPTTTKSA 229
          ||  :  ||      ||| :  ||      |||     ||     |||     ||| :
Db    241 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 300

Qy    230 PTTPKEPAPTTPKKPAPTTPKEPAPTTPKEPTPTTPKEPAPTTKEPAPTTPKEPAPTAPK 289
```



RESULT 6

US-08-928-361B-6
; Sequence 6, Application US/08928361B
; Patent No. 6071518
; GENERAL INFORMATION:
; APPLICANT: Petersen, Carolyn
; TITLE OF INVENTION: PEPTIDES, POLYPEPTIDES, GLYCOPROTEINS,
; TITLE OF INVENTION: THEIR FUNCTIONAL MUTANTS, VARIANTS, ANALOGS AND FRAGMENTS
; TITLE OF INVENTION: FOR TREATMENT AND DETECTION/DIAGNOSIS OF CRYPTOSPORIDIUM
; TITLE OF INVENTION: SPECIES INFECTIONS
; NUMBER OF SEQUENCES: 30
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: PETERS, VERNY, JONES & BIKSA
; STREET: 385 Sherman Avenue, Suite 6
; CITY: Palo Alto
; STATE: CA
; COUNTRY: USA
; ZIP: 94306-1840
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/928,361B
; FILING DATE: 12-SEP-1997
; CLASSIFICATION:
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 60/026,062
; FILING DATE: 13-SEP-1996
; ATTORNEY/AGENT INFORMATION:
; NAME: Verny, Hana
; REGISTRATION NUMBER: 30,518
; REFERENCE/DOCKET NUMBER: 480.76-1(HV)
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 650-324-1677
; TELEFAX: 650-324-1678
; INFORMATION FOR SEQ ID NO: 6:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 1721 amino acids
; TYPE: amino acid
; STRANDEDNESS:
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-928-361B-6

Query Match 8.6%; Score 489; DB 3; Length 1721;
Best Local Similarity 22.1%; Pred. No. 2.4e-24;
Matches 284; Conservative 82; Mismatches 476; Indels 442; Gaps 41;



```
Qy    350 PKE-------------------------------------------PSPTT------- 357
Db    407 TSETESVIKPDEWCWLEKNGECEAKGATYVGVIGKDGRIENGMAFTMIPNDDTHVRFRFK 466

Qy    358 ---------------------------TKEP--------------------APTTPKE 368
Db    467 VKDVGNTISVRCRKGAGKLEFPDRSLDFTIPPVAGHNSCSIIVGVSGDGKIHVSPYGSKD 526

Qy    369 PAPTTPKKPAPTTPKE-----------------PAPTTPKEPAPTTTKKPAPTAPKEPA 410
Db    527 VSLIS----APIQPSELFNEVYCDTCTAKYGAIHSGYQTSADFVTTTTAKPTTTT--TGA 580

Qy    411 PTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPTTP 470
Db    581 PGQPTTTTTGSPSKPTTTTTTKATTTT------TTLNPIITTTTQKPTTTT------TT- 627

Qy    471 KAAAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEPAPTTPKKPA 530
Db    628 --KVPGKPPIATTTTTLKPIVTTTTTKATTTTTTTVPTT-----TTTTKRDEMTTTTTPL 680

Qy    531 PK----ELAPTTTKEPTS--------------TTSDKPAPTTPKG-TAPTTPKEPAPTT 570
Db    681 PDIGDIEITPIPIEKMLDKYTRMIYDYNSGLLLDSNDEPIPGSQAGQIADTSNLFPVQTH 740

Qy    571 PKEPAPTTPKGTAPTTLKE---PAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKET 627
Db    741 KSTGLPIDPMVGLPFDPKSGNLVHPYTNQTMSGLSVSYLAAKNLTVDTDETYG--LPIDT 798

Qy    628 APTTPKEPAPTTPKKPAPTTPETPPPTTSEVSTPT--------TTKEPTTIHKS------ 673
Db    799 LTGYPLDPVSLIPFN--PETGELFDPISDEIMNGTIAGIVSGISASESLLSQKSALIDPA 856

Qy    674 -------------------------PDEST-------------PELSAEPTPKALENSPK 695
Db    857 TNMVVGEFGGLLNPATGVMIPGFLGPSEQTQFSPEIEDGGIIPPEVAAANADKFKLSIP- 915

Qy    696 EPGVPTTKTPAATKPEMTTTAKDKTTER-----------------DLRTTPETTTAAPKM 738
Db    916 -PSVP------ESIPE-----KDQKIDSISELMYDIESGRLIGQVSKRPIPGSIAGDLNP 963

Qy    739 TKETATTTEKTTESKITATT-TQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTE 797
Db    964 IMKTPTQTDSVTGKPIDPTTGLPFNPPTGHLINPTNNNTMDSSFAGAYKYAVSNGIKTDN 1023

Qy    798 IMNKP--EETAKPKD-------RATNSKATTPKPQKPTKAPKKPTSTKKPKTMP------ 842
Db   1024 VYGLPVGEITGLPKDPGSDIPFNSTTGELVDPSTGKPINNSTAGIVSGKPGLPPIEDENG 1083

Qy    843 ---------------RVRKPKTTPTPRKMTS--TMPE---------------------- 862
Db   1084 NLFDPSTNLPIDGNNQLVNPETNSTVSGSTSGTTKPKPGIPVNGGGVVPDEEAKDQADKG 1143

Qy    863 -----LNPTSRIAEAMLQTTTRPNQTPNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMP 917
Db   1144 KDGLIVPPTNSINKDPVTNTQYSNTTGNI----INP--ETGKVIPGSLPGSLNYPSFNTP 1197

Qy    918 EVTPD--------MDYLPRVPNQGIIINP--------------MLSDETNICNGK----- 950
Db   1198 QQTDEITGKPVDTVTGLPYDPSTGEIIDPATKLPIPGSVAGDEILTEVLNITTDEVTGLP 1257

Qy    951 ----------PVDGLTTLRNGTLV 964
Db   1258 IDLETGLPRDPVSGLPQLPNGTLV 1281
```

RESULT 7
US-07-638-431-2
; Sequence 2, Application US/07638431
; Patent No. 5198535
; GENERAL INFORMATION:
; APPLICANT: Hoffman, Stephen L.
; APPLICANT: Charoenvit, Yupin
; APPLICANT: Hedstrom, Richard
; APPLICANT: Khusmith, Srisin
; APPLICANT: Rogers IV, William O.
; TITLE OF INVENTION: Protective malaria sporozoite surface protein
; TITLE OF INVENTION: immunogen and gene
; NUMBER OF SEQUENCES: 2
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: A. David Spevack
; STREET: NMRDC Building 1 T-12 National Naval
; STREET: Medical Center
; CITY: Bethesda
; STATE: MD
; COUNTRY: USA
; ZIP: 20814-5044
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.24
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/07/638,431
; FILING DATE: 19910110
; CLASSIFICATION: 424
; ATTORNEY/AGENT INFORMATION:
; NAME: Spevack, Avrom D.
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (301) 295-6759
; TELEFAX: (301) 295-4033
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 826 amino acids
; TYPE: AMINO ACID
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-07-638-431-2

Query Match 8.6%; Score 488.5; DB 1; Length 826;
Best Local Similarity 26.4%; Pred. No. 1.1e-24;
Matches 153; Conservative 54; Mismatches 223; Indels 149; Gaps 34;

```
Qy    168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNPNN 332

Qy    221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649
```

```
Qy    631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
          |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    685 PTPKALENSPK-------------------EPGVPTTK 703
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   8
PCT-US92-00018-2
; Sequence 2, Application PC/TUS9200018
;  GENERAL INFORMATION:
;    APPLICANT:  Hoffman, Stephen L.
;    APPLICANT:  Charoenvit, Yupin
;    APPLICANT:  Hedstrom, Richard
;    APPLICANT:  Khusmith, Srisin
;    APPLICANT:  Rogers IV, William O.
;    TITLE OF INVENTION:  Protective malaria sporozoite surface protein
;    TITLE OF INVENTION:  immunogen and gene encoding
;    NUMBER OF SEQUENCES:  2
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  A. David Spevack
;      STREET:  NMRDC Building 1 T-12 National Naval
;      STREET:  Medical Center
;      CITY:  Bethesda
;      STATE:  MD
;      COUNTRY:  USA
;      ZIP:  20814-5044
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.24
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  PCT/US92/00018
;      FILING DATE:  19920103
;      CLASSIFICATION:  424
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Spevack, Avram D.
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (301) 295-6759
;      TELEFAX:  (301) 295-4033
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  826 amino acids
;      TYPE:  AMINO ACID
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  protein
PCT-US92-00018-2


  Query Match             8.6%;  Score 488.5;  DB 5;  Length 826;
  Best Local Similarity   26.4%;  Pred. No. 1.1e-24;
  Matches  153;  Conservative   54;  Mismatches  223;  Indels  149;  Gaps   34;

Qy    168 TPTTIKSAPTTPKEPA-PTTTKSAPTTPKEPA-PTTTKEP----APTTPKEP-APTTTKE 220
          ||  ::  |  |  |  |       |: |:|| |    :|      |  |  |  |
Db    273 TPCKVRDCPQIPIPPVIPNKIPEKPSNPEEPVNPNDPNDPNNPNNPNNPNNPNNPNNN 332

Qy    221 PAPTTTKSAPTTPKEP-APTTPKKP-APTTPKEP-APTTPKEP-TPTTPKEPAPTTKEPA 276
          |      : |  |  |  |  |  |  |  |  |  |  |  |  |  |  | :|:      |
Db    333 PNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNNPNDPSNPNNHPK 392

Qy    277 PTTPKEPAPTAPK--KPAPTTPKEPAPTTPKEPAPTTTKEPSPTTPKEPAPTTTKSAPTT 334
            ||   |  ||  || |  |    |: | :| |       |: | :| |    : |:
Db    393 RRNPKRRNPNKPKPNKPNPNKPNPNEPSNPNKPNPN-----EPSNPNKPNP----NEPSN 443

Qy    335 TKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPA-PTTPKKP-APTTPKEPAPTTPKE 392
            :| |    : |: | :|:|       |  | ||: |  |  | ||: | |  |: | |
Db    444 PNKPNP----NEPSNPNKPNPN-----EPLNPNEPSNPNEPSNPNAPSNPNE--PSNPNE 492

Qy    393 PAPTTTKKPAPTAPKEP-APTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAP 451
          |:        |  |  |  |: | |  |: |||                    |: | |  |
Db    493 PS-------NPNEPSNPNEPSNPNE--PSNPKK-----------------PSNPNE--P 523

Qy    452 TTPEEPTPTTPEEPAPTTPKAAAPNTPKEP-APTTPKEPAPTTPKEPAPTTPKETAPTTP 510
          : | |  |  | ||       : || |  |  |: |:|  |: |||  |: | |
Db    524 SNPNE--PLNPNEP-------SNPNEPSNPNEPSNPEE--PSNPKE--PSNPNE------ 564

Qy    511 KGTAPTTLKEPAPTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPTTPKEPAPTT 570
                      |: |::| |:|  |:  |||    |:   |  |:   |  | |  :
Db    565 -----------PSNPEEPNPEE--PSNPKEP----SNPEEPINPEELNPKEPSNPEESN 606

Qy    571 PKEPAPTTPKGTAPTTLKEPAPTTPKKPAPKELAPTTTKGPTSTTSDKPAPTTPKETAPT 630
          ||||                |  : ||:|    : |   : |    :   |      |
Db    607 PKEPI------------NPEESNPKEP----INPEDNENPLIIQDEPIEPRNDSNVIPI 649

Qy    631 TPKEP------APTTPKKPAPTTPETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAE 684
          |  |          |: |: :  | | |  :  ::  | |    |   |:|  |
Db    650 LPIIPQKGNNIPSNLPENPSDSEVEYPRPNDNGENSNNTMKSKKNI---PNEPIPSPGDN 706

Qy    685 PTPKALENSPK-------------------EPGVPTTK 703
          |     |  ||                    || :|  :
Db    707 PYKGHEERIPKPHRSNDDYVYDNNVNKNNKDEPEIPNNE 745


RESULT   9
5202236-13
;Patent No. 5202236
;    APPLICANT: MAUGH, KATHY J.;ANDERSON, DAVID M.;STRAUSBERG,
;SUSAN L.;MCCANDLISS, RUSS;WEI, TENA;FILPULA, DAVID
;    TITLE OF INVENTION: METHOD OF PRODUCING BIOADHESIVE
;PROTEIN
;    NUMBER OF SEQUENCES: 39
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER: US/07/528,762
;      FILING DATE: 25-MAY-1990
;      APPLICATION NUMBER: 82,456
;      FILING DATE: 07-AUG-1987
;      APPLICATION NUMBER: 933,945
;      FILING DATE: 24-NOV-1986
;      APPLICATION NUMBER: 650,128
;      FILING DATE: 13-SEP-1984
;SEQ ID NO:13:
;      LENGTH: 652
5202236-13


  Query Match             8.4%;  Score 476.5;  DB 6;  Length 652;
  Best Local Similarity   29.7%;  Pred. No. 5.3e-24;
  Matches  221;  Conservative   71;  Mismatches  284;  Indels  169;  Gaps   45;

Qy     51 PKITTAKPINPRPSLPPNSDTSKETSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKE 110
          ||:|      |:|| ||                | ::| |           | | |
Db     11 PKMTYPPTYKPKPSYPP---------------TYKSKPTY----------KPKIT---- 40

Qy    111 TQSIEKTSAKDLAPTSKVLAKPTPKAETTTKGPALTTPKEPTPTTPKE----PASTTPKE 166
                  || |  |||:          |    ||:  | | |     | :  ||
Db     41 ----------YPPTYK--AKPS--------YPPTYKPKKTYPPTYKPKLTYPPTYKPKP 79

Qy    167 PTPTTIKSAPT-TPKEPAPTTTKSAPTTPKEPAPTTTKEPAPTTPKEPAPTTTKEPA--P 223
            | | || ||  ||   | | |: |: |     |  |  |    ||   | | |  | |
Db     80 SYPPTYKSKPTYKPKITYPPTYKAKPSYPPTYKPKKTYPPT-YKPKLTYPPTYKPKASYP 138

Qy    224 TTTKSAPTTPKEPAPTTPKKPAPT-TPKEPAPTT----PKEPTPTTPKEPAPTTKEPAPT 278
            | |  |: |  |:  | |   ||  ||   | |    |  |    | :  | | :|  |
Db    139 PTYKPKPSYP--PSYKTKKTYPPTYKPKLTYPPTYKPKPSYPPSYKPNKTYPPTYKPKLT 196

Qy    279 TPKEPAPTAPKKPA--PTTPKEPA--PTTPKEPAPTTTKEPSPTTP-----KEPAPTTTK 329
          |    ||   ||:  ||   :|:  ||   :|:   | :  |: |      |   | | |
Db    197 YP----PTYKAKPSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYPSLIKAKPSYPPTYK 252

Qy    330 SAPTTTKEPA-------PTTTKSAPTTPKEPSPTTTKEPA--PTTPKEPA----PTTPKK 376
          : ||   :|        |:| |: || |    ||   :|:  ||   :|     ||   |
Db    253 AKPTYKAKPTYKAKPTYPSTYKAKPTYP----PTYKAKPSYPPTYKAKPTYKAKPTYKAK 308
```

```
Qy    377 PA--PTTPKEPA--PTTPKEPAPTTTKKPAPTAPKEPAPTTPKETAPTTPKKLTPTTPEK 432
          |:  ||   :|:  ||   :|:   | |  |: |    ||   : :     |: ||   |
Db    309 PSYPPTYKAKPSYPPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKVKPT--YK 362

Qy    433 LAPTTPE--KPAPTTPEELAPTTPEEPT--PTTPEEPA-PTTPKAAAPNTPKEPAPTTPK 487
            || |   |  |: |    ||   :|:  ||   :|: | | | | |: |    ||
Db    363 AKPTYPSTYKAKPSYP----PTYKAKPSYPPTYKAKPSYPPTYK-AKPSYP----PTYKA 413

Qy    488 EPA-PTT----PKEPAPTTPKETAPTT--PKGTAPTTLKEPAPTTPKKPAPKELAPTTTK 540
          :|  |:|    |  |    || : | |   | : |:| |  :   |   |     ||
Db    414 KPTYPSTYKAKPSYPPTYKPKISYPPTYKAKPSYPSTYKAKSSYPPTYKAKPSYPPTYKA 473

Qy    541 EPTSTTSDKPAPTTPKGTAPTTPKEPAPTTPKEPAPTTPKGTAPTTLKEPA--PTTPKKP 598
          :||  :: |  ||      |: |    ||   :|:        ||   :|:  ||   ||
Db    474 KPTYPSTYKAKPTYK--AKPSYP----PTYKAKPS------YPPTYKAKPSYPPTYKAKP 521

Qy    599 APKELAPTTTKGPTSTTSDKPAPTTPKETAPTTPKEPA--PTTPKKPA-PTT----PETP 651
            |   ||    |:   : |  |: |    ||   :|:  ||   ||  |:|    |  |
Db    522 TYK-AKPTYKAKPSYPPTYKAKPSYP----PTYKAKPSYPPTYKAKPTYPSTYKAKPSYP 576

Qy    652 PPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAATKPE 711
          |   :: | | |       :|:     |  |:|| ||     :|  |:|      ||
Db    577 PTYKAKPSYPPT-------YKAKPSYPPTYKAKPTYKA------KPTYPST---YKAKPS 620

Qy    712 MTTTAKDKTTERDLRTTPETTTAAP 736
            | | | |:   : | |  | |
Db    621 YPPTHKAKPTYKAKPSYPPTHKAKP 645


RESULT  10
US-09-219-849-5
; Sequence 5, Application US/09219849
; Patent No. 6150081
; GENERAL INFORMATION:
;  APPLICANT:  VAN HEERDE, GEORGE V.
;  APPLICANT:  VAN RIJN, ALEXIS C.
;  APPLICANT:  BOUWSTRA, JAN B.
;  APPLICANT:  DE WOLF, FREDERIK A.
;  APPLICANT:  MOOBROEK, ANDREAS
;  APPLICANT:  WERTEN, MARC W.T.
;  APPLICANT:  WIND, RICHELE D.
;  APPLICANT:  VAN DEN BOSCH, TANJA J.
;  TITLE OF INVENTION:  SILVER HALIDE EMULSIONS WITH RECOMBINANT COLLAGEN
;  TITLE OF INVENTION:  SUITABLE FOR PHOTOGRAPHIC APPLICATION AND ALSO THE
;  TITLE OF INVENTION:  PREPARATION THEREOF
;  FILE REFERENCE:  2728-2
;  CURRENT APPLICATION NUMBER: US/09/219,849
;  CURRENT FILING DATE:  1998-12-23
;  NUMBER OF SEQ ID NOS:  50
;  SOFTWARE:  PatentIn Ver. 2.1
; SEQ ID NO 5
;   LENGTH: 960
;   TYPE: PRT
;   ORGANISM: Artificial Sequence
;   FEATURE:
;   OTHER INFORMATION: Description of Artificial Sequence: Illustrative
;   OTHER INFORMATION: amino acid sequence
US-09-219-849-5


  Query Match             7.9%;  Score 452;  DB 4;  Length 960;
  Best Local Similarity   25.6%;  Pred. No. 3.3e-22;
  Matches  173;  Conservative   60;  Mismatches  309;  Indels  134;  Gaps   41;

Qy    132 PTPKAETTTKGPALTTPKEPTPTTPKEPASTTPKEPTPTTIKSAPTTPKEPAP--TTTKS 189
            | |     :: |    | | | |  | | |  |       : |   ::| |      :
Db     72 PGPAGPPGSRDPG--PPGAPGPAGP--PGSRDPGPPGAPG-PAGPPGSRDPGPPGAPGPA 126

Qy    190 APTTPKEPAPTTTKEPA-PTTPKEPAPTTTKEPAPTTTKSAPTTPKEPAPTTP---KKPA 245
            |   ::| |     || |   ::| |     ||        | |  | |  |   : |
Db    127 GPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPG 186

Qy    246 PTTPKEPAPTTP---KEPTPTTPKEPA--PTTKEPAPTTPKEPA--------------- 284
          | | | | |   ::| |      ||  | :::| |     ||
Db    187 P--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAHGPAGPKGAHGPAGPKGAHG 244

Qy    285 ---------PTAPK-KPAPTTP---KEPAPTTPKEPA-PTTTKEPSPTTPKEPAPTTTKS 330
                  | ||  | | |   ::| |     || |  :::| |  |  | |    :
Db    245 PAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGAPGP----A 298

Qy    331 APTTTKEPAPTTTKSAPTTPKEPSPTTTKEPAPTTPKEPAPTTP---KKPAPTTPKEPAP 387
            |  :::| |       | |     |  :::| | |  | | |  |   : | | |  | |
Db    299 GPPGSRDPGP------PGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGP--PGAPGP 348

Qy    388 TTP---KEPAPTTTKKPA-------PTAPKEPAPTTPKETAPTTPKKLTPTTPEKLAPTT 437
            |   ::| |     ||        |  | | | |  |  :    |   | |    |
Db    349 AGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP----PGAPGPAGPPG 404

Qy    438 PEKPAPT-TPEELAPTTPEEPTPTTPEEPAPTTPKAAA-PNTPKEPAPTTP---KEPAPT 492
            | |   |    |   :| | |  | | |  |  :  |  | |  | |   ::| |
Db    405 SRDPGPPGAPGPAGPPGSRDPGP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPP 462

Qy    493 TPKEPA-------PTTPKET-APTTPKGT--------APTTLKEPAPTTPKKP-APKELA 535
            ||       |  ||    |  |||          ||     |   |  | ||
Db    463 GAHGPAGPKGAHGPAGPKGAHGPAGPKGAHGPAGPKGAPGPAGPPGSRDPGPPGAPGPAG 522

Qy    536 PTTTKEPTSTTSDKPAPTTPKGTA-PTTPKEPAPTTP---KEPAPTTPKGTA-PTTLKEP 590
            |  :::|    :  | |  | |:  |  | | | |  ::| |    | | |   ::|
Db    523 PPGSRDPGPPGA--PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDP 580

Qy    591 APTTPKKPAP------KELAPTTTKGPTST--TSDKPAPTTPKETAPTTPKEPAPTTPKK 642
            |  | | |       ::  |    ||      : |   | |    |   ::| |  |
Db    581 GP--PGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP--PGA 636

Qy    643 PAPTTP----ETPPPTTSEVSTPTTTKEPTTIHKSPDESTPELSAEPTPKALEN-----S 693
          | |  |    :  ||     : |  :::|     :|  : |  | :| |
Db    637 PGPAGPPGSRDPGPPGAPGPAGPPGSRDPGP-PGAPGPAGPPGSRDPGPPGAPGPAGPPG 695

Qy    694 PKEPGVPTTKTPAATK 709
            ::|| |    ||  |
Db    696 SRDPGPPGAHGPAGPK 711


RESULT  11
US-08-479-537A-5
; Sequence 5, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
```

```
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  5:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  1867 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1727
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-5


  Query Match             7.4%;  Score 424.5;  DB 2;  Length 1867;
  Best Local Similarity   23.7%;  Pred. No. 4.5e-20;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
          |    :||:|    |    ||               | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
             :||| :|.       |||     |             ||      | | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|        |||     | | |       ||      | | |     |:

Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             : | | :  |          ||      |  |||        ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |         |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||      |  | |       ||      |  | |         ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |      |:     ||      |  | |     |:  :  |     |  |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
            | |     :||:    |  | |     |:  :  |  |  |||      | ||     |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
            | |           :| |   ||       ||    :| |  | |     |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
            |  | |      :||      |  | |           :| |   ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |      :||      |  | |           :| | :  | |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :          :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                |       |  |  ||     :  :       |  ||      |::    | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
            : ||      |  |    |      | |     :||  |:  |              | |
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy     884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
            | :  |   |      |:     |           ||   |  |   :      |
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy     943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
            |:  : :|  | |    :|   |           |   | |         |: |  |
Db     958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  12
US-09-083-116-5
; Sequence 5, Application US/09083116
; Patent No. 6203795
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
```

```
;       STREET:  P.O. Box 1404
;       CITY:  Alexandria
;       STATE:  Virginia
;       COUNTRY:  United States
;       ZIP:  22313-1404
;     COMPUTER READABLE FORM:
;       MEDIUM TYPE:  Floppy disk
;       COMPUTER:  IBM PC compatible
;       OPERATING SYSTEM:  PC-DOS/MS-DOS
;       SOFTWARE:  PatentIn Release #1.0, Version #1.30
;     CURRENT APPLICATION DATA:
;       APPLICATION NUMBER:  US/09/083,116
;       FILING DATE:
;       CLASSIFICATION:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  08/479,537
;       FILING DATE:
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  WO PCT/FR91/00835
;       FILING DATE:  23-OCT-1991
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/039,320
;       FILING DATE:  04-APR-1993
;     PRIOR APPLICATION DATA:
;       APPLICATION NUMBER:  US 08/403,576
;       FILING DATE:  14-MAR-1995
;     ATTORNEY/AGENT INFORMATION:
;       NAME:  Teskin, Robin L.
;       REGISTRATION NUMBER:  35,030
;       REFERENCE/DOCKET NUMBER:  017753-025
;     TELECOMMUNICATION INFORMATION:
;       TELEPHONE:  (703) 836-6620
;       TELEFAX:  (703) 836-2021
;   INFORMATION FOR SEQ ID NO:  5:
;     SEQUENCE CHARACTERISTICS:
;       LENGTH:  1867 amino acids
;       TYPE:  amino acid
;       STRANDEDNESS:  single
;       TOPOLOGY:  linear
;     MOLECULE TYPE:  peptide
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  128..1727
;       OTHER INFORMATION:  /note= "The amino acids spanning
;       OTHER INFORMATION:  128 to 1727 constitute a repeated region wherein the repeat i
;       OTHER INFORMATION:  20 amino acids, 17 of which are fixed.  The number of such
;       OTHER INFORMATION:  repeats varies from 1 to 40."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  134
;       OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  144
;       OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;       OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC, ACA
;       OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  147
;       OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;       OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC, CCA
;       OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;     FEATURE:
;       NAME/KEY:  Peptide
;       LOCATION:  1..21
;       OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;       OTHER INFORMATION:  21 amino acid precursor sequence."
US-09-083-116-5
```

```
 Query Match             7.4%;  Score 424.5;  DB 4;  Length 1867;
 Best Local Similarity   23.7%;  Pred. No. 4.5e-20;
 Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||           | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |     ||||      |  |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP-----------APTTTKEPAPTTPKEPAPTT 217
             :||| :|         |||      |                ||     |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|         |||      |  | |         ||     |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |         ||      |  |||         ||     |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |          |||       |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||      |  | |         ||      |  | |         ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |      |:    ||      |  | |      |:  :  |      |  |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
           | |     :||:    |  | |      |:  :  |  |  |||      | ||      |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||      ||     :| |  | |      |:    ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
           |  | |        :||      |  | |           :| |   ||      | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |      | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||      |  |     : ::|      | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
            :|   :||       |:  | |  | :|  ||      |:    :             |||
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
                |       |  |  ||      :  :    |  ||      |::   | |
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
           : ||      |  |   |       | |    :||  |:  |                  | |
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy     884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
            | :  |   |       |:      |              ||   |  |   :      |
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy     943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
```

```
          |:  : :|  | |    :|   |          |   | |        |: |  |
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003


RESULT  13
US-08-479-537A-2
; Sequence 2, Application US/08479537A
; Patent No. 5861381
;  GENERAL INFORMATION:
;    APPLICANT:  CHAMBON, Pierre
;    APPLICANT:  KIENY, Marie-Paule
;    APPLICANT:  LATHE, Richard
;    APPLICANT:  HAREUVENI, Mara
;    TITLE OF INVENTION:  PHARMACEUTICAL COMPOSITION FOR THE
;    TITLE OF INVENTION:  TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
;    NUMBER OF SEQUENCES:  5
;    CORRESPONDENCE ADDRESS:
;      ADDRESSEE:  BURNS, DOANE, SWECKER & MATHIS, L.L.P.
;      STREET:  P.O. Box 1404
;      CITY:  Alexandria
;      STATE:  Virginia
;      COUNTRY:  United States
;      ZIP:  22313-1404
;    COMPUTER READABLE FORM:
;      MEDIUM TYPE:  Floppy disk
;      COMPUTER:  IBM PC compatible
;      OPERATING SYSTEM:  PC-DOS/MS-DOS
;      SOFTWARE:  PatentIn Release #1.0, Version #1.30
;    CURRENT APPLICATION DATA:
;      APPLICATION NUMBER:  US/08/479,537A
;      FILING DATE:  07-JUN-1995
;      CLASSIFICATION:  514
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  FR 90/13101
;      FILING DATE:  23-OCT-1990
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  WO PCT/FR91/00835
;      FILING DATE:  23-OCT-1991
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/039,320
;      FILING DATE:  04-APR-1993
;    PRIOR APPLICATION DATA:
;      APPLICATION NUMBER:  US 08/403,576
;      FILING DATE:  14-MAR-1995
;    ATTORNEY/AGENT INFORMATION:
;      NAME:  Teskin, Robin L.
;      REGISTRATION NUMBER:  35,030
;      REFERENCE/DOCKET NUMBER:  017753-025
;    TELECOMMUNICATION INFORMATION:
;      TELEPHONE:  (703) 836-6620
;      TELEFAX:  (703) 836-2021
;  INFORMATION FOR SEQ ID NO:  2:
;    SEQUENCE CHARACTERISTICS:
;      LENGTH:  2035 amino acids
;      TYPE:  amino acid
;      STRANDEDNESS:  single
;      TOPOLOGY:  linear
;    MOLECULE TYPE:  peptide
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  128..1899
;      OTHER INFORMATION:  /note= "The amino acids spanning
;      OTHER INFORMATION:  128 to 1899 constitute a repeated region wherein the repeat i
;      OTHER INFORMATION:  20 amino acids, 17 of which are fixed. The number of such
;      OTHER INFORMATION:  repeats varies from 1 to 40."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  134
;      OTHER INFORMATION:  /note= "Amino acid 134 is X1 = Xaa
;      OTHER INFORMATION:  Xaa Xaa which is the codon for Pro or Ala wherein Pro = CCT,
;      OTHER INFORMATION:  CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  144
;      OTHER INFORMATION:  /note= "Amino acid 144 is Y = Xaa
;      OTHER INFORMATION:  which is the codon for Thr or Asn wherein Thr = ACT, ACC,
;      OTHER INFORMATION:  or ACG; and Asn = AAT or AAC."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  147
;      OTHER INFORMATION:  /note= "Amino acid 147 is X2 = Xaa
;      OTHER INFORMATION:  which is the codon for Pro or Ala wherein Pro = CCT, CCC,
;      OTHER INFORMATION:  or CCG; and Ala = GCT, GCC, GCA, or GCG."
;    FEATURE:
;      NAME/KEY:  Peptide
;      LOCATION:  1..21
;      OTHER INFORMATION:  /note= "Amino acids 1 to 21 are a
;      OTHER INFORMATION:  21 amino acid precursor sequence."
US-08-479-537A-2


  Query Match             7.4%;  Score 424.5;  DB 2;  Length 2035;
  Best Local Similarity   23.7%;  Pred. No. 5e-20;
  Matches  255;  Conservative   69;  Mismatches  511;  Indels  241;  Gaps   53;

Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
           | |||          | :   :|| | || ||| :  |:  :: |: :: || ||:  :
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP--------------TPTTPKEPASTTPKEPTPTTIKSAP 176
           |    :||:|    |    ||                | |     ||||     |   |||
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
           :||| :|         |||      |              ||      |  | |     |:
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
             :  | | :|         |||      |  | |          ||      |  | |     |:
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
             :  | | :  |         ||      |  |||        ||      |  | |     |:
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
             :  |  |   ||       |||      |  |          |||      |  |
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
            ||     |  | |        ||      |  | |          ||      |  |||     |
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
           :  :  |    | |      |:     ||      |  | |      |:  :  |      | |
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
            | |     :||:     |  | |      |:  :  |  |  |||      | ||      |
Db     472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy     524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
           | |           :| |   ||       ||    :| |  | |       |:     ||
Db     531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy     572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
             |  | |       :||      |  | |           :| |   ||    | ||
Db     586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy     618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
           :| |  | |       :||      |  | |           :| | :  | |       | |
Db     645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy     662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
             | :       :  :||      |  |    : ::|     | :  | |
Db     703 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762
```

```
Qy     708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
Db     763 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 820

Qy     764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
Db     821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy     824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
Db     869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX----------PGST 911

Qy     884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
Db     912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy     943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
Db     958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

RESULT 14
US-09-083-116-2
; Sequence 2, Application US/09083116
; Patent No. 6203795
; GENERAL INFORMATION:
; APPLICANT: CHAMBON, Pierre
; APPLICANT: KIENY, Marie-Paule
; APPLICANT: LATHE, Richard
; APPLICANT: HAREUVENI, Mara
; TITLE OF INVENTION: PHARMACEUTICAL COMPOSITION FOR THE
; TITLE OF INVENTION: TREATMENT OR PREVENTION OF A MALIGNANT TUMOR
; NUMBER OF SEQUENCES: 5
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: BURNS, DOANE, SWECKER & MATHIS, L.L.P.
; STREET: P.O. Box 1404
; CITY: Alexandria
; STATE: Virginia
; COUNTRY: United States
; ZIP: 22313-1404
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.30
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/09/083,116
; FILING DATE:
; CLASSIFICATION:
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: 08/479,537
; FILING DATE:
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: WO PCT/FR91/00835
; FILING DATE: 23-OCT-1991
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 08/039,320
; FILING DATE: 04-APR-1993
; PRIOR APPLICATION DATA:
; APPLICATION NUMBER: US 08/403,576
; FILING DATE: 14-MAR-1995
; ATTORNEY/AGENT INFORMATION:
; NAME: Teskin, Robin L.
; REGISTRATION NUMBER: 35,030
; REFERENCE/DOCKET NUMBER: 017753-025
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: (703) 836-6620
; TELEFAX: (703) 836-2021
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 2035 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: peptide
; FEATURE:
; NAME/KEY: Peptide
; LOCATION: 128..1899
; OTHER INFORMATION: /note= "The amino acids spanning
; OTHER INFORMATION: 128 to 1899 constitute a repeated region wherein the repea
; OTHER INFORMATION: 20 amino acids, 17 of which are fixed. The number of such
; OTHER INFORMATION: repeats varies from 1 to 40."
; FEATURE:
; NAME/KEY: Peptide
; LOCATION: 134
; OTHER INFORMATION: /note= "Amino acid 134 is X1 = Xaa
; OTHER INFORMATION: Xaa Xaa which is the codon for Pro or Ala wherein Pro = CC
; OTHER INFORMATION: CCC, CCA, or CCG; and Ala = GCT, GCC, GCA, or GCG."
; FEATURE:
; NAME/KEY: Peptide
; LOCATION: 144
; OTHER INFORMATION: /note= "Amino acid 144 is Y = Xaa
; OTHER INFORMATION: which is the codon for Thr or Asn wherein Thr = ACT, ACC,
; OTHER INFORMATION: or ACG; and Asn = AAT or AAC."
; FEATURE:
; NAME/KEY: Peptide
; LOCATION: 147
; OTHER INFORMATION: /note= "Amino acid 147 is X2 = Xaa
; OTHER INFORMATION: which is the codon for Pro or Ala wherein Pro = CCT, CCC,
; OTHER INFORMATION: or CCG; and Ala = GCT, GCC, GCA, or GCG."
; FEATURE:
; NAME/KEY: Peptide
; LOCATION: 1..21
; OTHER INFORMATION: /note= "Amino acids 1 to 21 are a
; OTHER INFORMATION: 21 amino acid precursor sequence."
US-09-083-116-2

Query Match 7.4%; Score 424.5; DB 4; Length 2035;
Best Local Similarity 23.7%; Pred. No. 5e-20;
Matches 255; Conservative 69; Mismatches 511; Indels 241; Gaps 53;

```
Qy      75 TSLTVNKETTVETKETTTTNKQTSTDGKEKTTSAKETQSIEKTSAKD-LAPTSKVLAKPT 133
Db      16 TVLTV----------VTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHS 65

Qy     134 P-KAETTTKGP--ALTTPKEP-------------TPTTPKEPASTTPKEPTPTTIKSAP 176
Db      66 PGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVT---SAP 122

Qy     177 TTPKEPAPTTT-------KSAPTTPKEP------------APTTTKEPAPTTPKEPAPTT 217
Db     123 --DNKPAPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 180

Qy     218 TKE--PAPTTT-------KSAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPKEPTPTT 264
Db     181 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 240

Qy     265 PKE--PAPTTKEP-------APTTPKEPAPTAPK----KPAPTTPKEPAPTTPKEPAPTT 311
Db     241 APDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTS 300

Qy     312 TKEPSPTTPKEPAPTT--TKSAPTTTKEPAPTT-----TKSAPTTPKEPSPTTTK----E 360
Db     301 APDXRP-XPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 359

Qy     361 PAPTTPKEPAPTTPK----KPAPTTPKEPAPTTPK----EPAPTTTKKPAPTAPKEPAPT 412
Db     360 SAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVT 419

Qy     413 TPKETAP----TTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEPAPT 468
Db     420 SAPDXRPXPGSTAPXAHGVTS----APDXRPXPGSTAPXAHGVTSAPDXRPX----PGST 471

Qy     469 TPKA----AAPNTPKEPAPTTPKEPAPTTPKEPAPTTPKETAPTTPKGT-APTTLKEPAP 523
```

```
Db    472 APXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRP-XPGSTAPXAHGVTSAPDXRPXPGS 530

Qy    524 TTP----------KKPAPKELAPTTTKEPTSTTSDKPAP--TTPKGTAPTTPKEPAPTTP 571
Db    531 TAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDXRPXPGSTAPXAHGVTS----APDXR 585

Qy    572 KEPAPTTPKG----TAPTTLKEPAPTTP----------KKPAPKELAPTTTKGPTSTTSD 617
Db    586 PXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAP-XAHGVTSAPDX 644

Qy    618 KPAP--TTPK----ETAPTTPKEPAPTTP----------KKPAPTTPETPPPTTSEVSTP 661
Db    645 RPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGS--TAPXAHGVTSAP 702

Qy    662 TTTKEPTT-------IHKSPDESTPELSAEPTPKALENSPKEPGVPTTKTPAA------- 707
Db    703 DXRPXPGSTAPXAHGVTSAPDXRPXPGST--APXAHGVTSAPDXRPXPGSTAPXAHGVTS 760

Qy    708 -TKPEMTTTAKDK---TTERDLRTTPETTTAAPKMTKETATTTEKTTESKITATTTQVTS 763
Db    763 DXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAPDXRPXPGSTAPXAHGVTSAP 762

Qy    764 TTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETAKPKDRATNSKATTPKPQ 823
Db    821 APDXRPXPG-----STAPXAHGVT------SAPDXRPXPGSTAPXAHGVTSAPDXRPXP- 868

Qy    824 KPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPTSRIAEAMLQTTTRPNQT 883
Db    869 -GSTAPXAHGVTSAPDXRP--XPGSTAPXAHGVTSA-PDXRPX------------PGST 911

Qy    884 -PNSKLVEVNPKSEDAGGAEGETPHMLLRPHVFMPEVTPDMDYLPRVPNQGIIINPMLSD 942
Db    912 APXAHGVTSAPDXRPXPGSTAPXAH----------GVTSAPDXRPXPGSTA----PXAHG 957

Qy    943 ETNICNGKPVDGLTT-LRNGTLVAFRGHYFWMLSPFSPPSPARRITEVWGIPSPID 997
Db    958 VTSAPDXRPXPGSTAPXAHGVTSA----------PDXRPXPGSTAPXAHGVTSAPD 1003
```

RESULT 15
US-08-276-967-2
; Sequence 2, Application US/08276967
; Patent No. 5851817
; GENERAL INFORMATION:
; APPLICANT: Hardy, Daniel M.
; APPLICANT: Garbers, David L.
; TITLE OF INVENTION: Species-Specific Egg-Binding Proteins of
; TITLE OF INVENTION: Sperm
; NUMBER OF SEQUENCES: 8
; CORRESPONDENCE ADDRESS:
; ADDRESSEE: Arnold, White & Durkee
; STREET: P. O. Box 4433
; CITY: Houston
; STATE: Texas
; COUNTRY: USA
; ZIP: 77210-4433
; COMPUTER READABLE FORM:
; MEDIUM TYPE: Floppy disk
; COMPUTER: IBM PC compatible
; OPERATING SYSTEM: PC-DOS/MS-DOS
; SOFTWARE: PatentIn Release #1.0, Version #1.25
; CURRENT APPLICATION DATA:
; APPLICATION NUMBER: US/08/276,967
; FILING DATE: Submitted Herewith
; CLASSIFICATION: 530
; ATTORNEY/AGENT INFORMATION:
; NAME: Kitchell, Barbara S.
; REGISTRATION NUMBER: 33,928
; REFERENCE/DOCKET NUMBER: UTSD:418\KIT
; TELECOMMUNICATION INFORMATION:
; TELEPHONE: 713-787-1400
; TELEFAX: 713-789-2679
; TELEX: 79-0924
; INFORMATION FOR SEQ ID NO: 2:
; SEQUENCE CHARACTERISTICS:
; LENGTH: 2476 amino acids
; TYPE: amino acid
; STRANDEDNESS: single
; TOPOLOGY: linear
; MOLECULE TYPE: protein
US-08-276-967-2

Query Match 7.4%; Score 419.5; DB 2; Length 2476;
Best Local Similarity 32.5%; Pred. No. 1.3e-19;
Matches 163; Conservative 43; Mismatches 177; Indels 119; Gaps 31;

```
Qy    406 PKEPAPTTPKETAPTTPKKLTPTTPEKLAPTTPEKPAPTTPEELAPTTPEEPTPTTPEEP 465
Db    313 PSETSVSTEKPVAPTE----KPTVPSEIYTIPTEKPMVHMEKPIVHT--EKPTVPT-EKP 365

Qy    466 APTTPKAAAPNTPKEPA----PTTPKEPAPTTPKEPAPTTPKETAPTTPKGTAPTTLKEP 521
Db    366 TIPTEKSTVPT--KKPTVFKEPTLPPE-GPTVPAE-RPTTPPEGPAVPPKG--PTVLTE- 418

Qy    522 APTTPKKPAPKELAPTTTKEPTSTTSDKPAPTTPKGTAPT-TPKEPAPTTPKEPAPTTPK 580
Db    419 ------------WPTSHTEKSTVHTEKPILPTGKSTIPTEKPMVPTKRT------TTP- 458

Qy    581 GTAPTTLKEPAPTTPKKPAPKELAPT-TTKGPTSTT---SDKPAPTTPKETAPTTPKEPA 636
Db    459 -TERTTIPAEKPTV---PIEKPMVPTERTTIPTERTTIPTEKPTVPTEKLTVPT--EKPI 512

Qy    637 PTTPKKPAPTTPETPPP-----TTSEVSTPT-----TTKEPTTIHKSPDESTPELSAEPT 686
Db    513 VPTEKPIVPTEKHTIPTEKLTVLTERTTTPTERTTIPTEKPTVPTEKPSVPT-EKPTVPT 571

Qy    687 PKALENSPKEPGVPTTKTPAATKPEMTTTAKDKTTERDLRTTPETTTAAPKMTKETATTT 746
Db    572 --------EEPTIPTEKLTVPT--ERTTTPTKRTTTPTIRTTTPTI--------RTTTPT 613

Qy    747 EKTTESKITATTTQVTSTTTQDTTPFKITTLKTTTLAPKVTTTKKTITTTEIMNKPEETA 806
Db    614 ERTT------TPTIRTTTPTERTT---IPTKKTT------VPTEKTIIPT------ERTI 652

Qy    807 KPKDRATNSKATTPKPQKPTKAPKKPTSTKKPKTMPRVRKPKTTPTPRKMTSTMPELNPT 866
Db    653 AP---------TTPQP-SPTLVPTQPAAVVMPST------SATTVTPRTTIASCP---PN 693

Qy    867 SRIAEAMLQTTTRPNQTPNSKL 888
Db    694 AHFERCACPVSCQ-SPTPNCEL 714
```

Search completed: April 26, 2002, 16:17:20
Job time: 231 sec